

## Regularized Learning for Domain Adaptation under Label Shifts

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, Animashree Anandkumar

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

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.

\*\*\*\*\*

## Generative Adversarial Models for Learning Private and Fair Representations

Chong Huang, Xiao Chen, Peter Kairouz, Lalitha Sankar, Ram Rajagopal

We present Generative Adversarial Privacy and Fairness (GAPF), a data-driven framework for learning private and fair representations of the data. GAPF leverages recent advances in adversarial learning to allow a data holder to learn "universal" representations that decouple a set of sensitive attributes from the rest of the dataset. Under GAPF, finding the optimal decorrelation scheme is formulated as a constrained minimax game between a generative decorrelator and an adversary. We show that for appropriately chosen adversarial loss functions, GAPF provides privacy guarantees against strong information-theoretic adversaries and enforces demographic parity. We also evaluate the performance of GAPF on multi-dimensional Gaussian mixture models and real datasets, and show how a designer can certify that representations learned under an adversary with a fixed architecture perform well against more complex adversaries.

\*\*\*\*\*

## Generative Adversarial Self-Imitation Learning

Junhyuk Oh, Yijie Guo, Satinder Singh, Honglak Lee

This paper explores a simple regularizer for reinforcement learning by proposing Generative Adversarial Self-Imitation Learning (GASIL), which encourages the agent to imitate past good trajectories via generative adversarial imitation learning framework. Instead of directly maximizing rewards, GASIL focuses on reproducing past good trajectories, which can potentially make long-term credit assignment easier when rewards are sparse and delayed. GASIL can be easily combined with any policy gradient objective by using GASIL as a learned reward shaping function. Our experimental results show that GASIL improves the performance of proximal policy optimization on 2D Point Mass and MuJoCo environments with delayed reward and stochastic dynamics.

\*\*\*\*\*

## INFORMATION MAXIMIZATION AUTO-ENCODING

Dejiao Zhang,Tianchen Zhao,Laura Balzano

We propose the Information Maximization Autoencoder (IMAE), an information theoretic approach to simultaneously learn continuous and discrete representations in an unsupervised setting. Unlike the Variational Autoencoder framework, IMAE starts from a stochastic encoder that seeks to map each input data to a hybrid discrete and continuous representation with the objective of maximizing the mutual information between the data and their representations. A decoder is included to approximate the posterior distribution of the data given their representations, where a high fidelity approximation can be achieved by leveraging the informative representations.

We show that the proposed objective is theoretically valid and provides a principled framework for understanding the tradeoffs regarding informativeness of each representation factor, disentanglement of representations, and decoding quality.

\*\*\*\*\*

Variadic Learning by Bayesian Nonparametric Deep Embedding

Kelsey R Allen,Hanul Shin,Evan Shelhamer,Josh B. Tenenbaum

Learning at small or large scales of data is addressed by two strong but divided frontiers: few-shot learning and standard supervised learning. Few-shot learning focuses on sample efficiency at small scale, while supervised learning focuses on accuracy at large scale. Ideally they could be reconciled for effective learning at any number of data points (shot) and number of classes (way). To span the full spectrum of shot and way, we frame the variadic learning regime of learning from any number of inputs. We approach variadic learning by meta-learning a novel multi-modal clustering model that connects bayesian nonparametrics and deep metric learning. Our bayesian nonparametric deep embedding (BANDE) method is optimized end-to-end with a single objective, and adaptively adjusts capacity to learn from variable amounts of supervision. We show that multi-modality is critical for learning complex classes such as Omniglot alphabets and carrying out unsupervised clustering. We explore variadic learning by measuring generalization across shot and way between meta-train and meta-test, show the first results for scaling from few-way, few-shot tasks to 1692-way Omniglot classification and 5k-shot CIFAR-10 classification, and find that nonparametric methods generalize better than parametric methods. On the standard few-shot learning benchmarks of Omniglot and mini-ImageNet, BANDE equals or improves on the state-of-the-art for semi-supervised classification.

\*\*\*\*\*

DL2: Training and Querying Neural Networks with Logic

Marc Fischer,Mislav Balunovic,Dana Drachslers-Cohen,Timon Gehr,Ce Zhang,Martin Vechev

We present DL2, a system for training and querying neural networks with logical constraints. The key idea is to translate these constraints into a differentiable loss with desirable mathematical properties and to then either train with this loss in an iterative manner or to use the loss for querying the network for inputs subject to the constraints. We empirically demonstrate that DL2 is effective in both training and querying scenarios, across a range of constraints and data sets.

\*\*\*\*\*

Local Image-to-Image Translation via Pixel-wise Highway Adaptive Instance Normalization

Wonwoong Cho,Seunghwan Choi,Junwoo Park,David Keetae Park,Tao Qin,Jaegul Choo

Recently, image-to-image translation has seen a significant success. Among many approaches, image translation based on an exemplar image, which contains the target style information, has been popular, owing to its capability to handle multi-modality as well as its suitability for practical use. However, most of the existing methods extract the style information from an entire exemplar and apply it to the entire input image, which introduces excessive image translation in irrelevant image regions. In response, this paper proposes a novel approach that jointly extracts out the local masks of the input image and the exemplar as targeted regions to be involved for image translation. In particular, the main novelty o

Our model lies in (1) co-segmentation networks for local mask generation and (2) the local mask-based highway adaptive instance normalization technique. We demonstrate the quantitative and the qualitative evaluation results to show the advantages of our proposed approach. Finally, the code is available at <https://github.com/AnonymousIclrAuthor/Highway-Adaptive-Instance-Normalization>

\*\*\*\*\*

The Limitations of Adversarial Training and the Blind-Spot Attack

Huan Zhang\*, Hongge Chen\*, Zhao Song, Duane Boning, Inderjit S. Dhillon, Cho-Jui Hsieh

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 networks (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 distribution 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

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.

\*\*\*\*\*

Inference of unobserved event streams with neural Hawkes particle smoothing

Hongyuan Mei, Guanghui Qin, Jason Eisner

Events that we observe in the world may be caused by other, unobserved events. We consider sequences of discrete events in continuous time. When only some of the events are observed, we propose particle smoothing to infer the missing events. Particle smoothing is an extension of particle filtering in which proposed events are conditioned on the future as well as the past. For our setting, we develop a novel proposal distribution that is a type of continuous-time bidirectional LSTM. We use the sampled particles in an approximate minimum Bayes risk decoder that outputs a single low-risk prediction of the missing events. We experiment

in multiple synthetic and real domains, modeling the complete sequences in each domain with a neural Hawkes process (Mei & Eisner, 2017). On held-out incomplete sequences, our method is effective at inferring the ground-truth unobserved events. In particular, particle smoothing consistently improves upon particle filtering, showing the benefit of training a bidirectional proposal distribution.

\*\*\*\*\*

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

Chien-Sheng Wu, Richard Socher, Caiming Xiong

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

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.

\*\*\*\*\*

#### A Better Baseline for Second Order Gradient Estimation in Stochastic Computation Graphs

Jingkai Mao, Jakob Foerster, Tim Rocktäschel, Gregory Farquhar, Maruan Al-Shedivat, Shimon Whiteson

Motivated by the need for higher order gradients in multi-agent reinforcement learning and meta-learning, this paper studies the construction of baselines for second order Monte Carlo gradient estimators in order to reduce the sample variance. Following the construction of a stochastic computation graph (SCG), the Infinitely Differentiable Monte-Carlo Estimator (DiCE) can generate correct estimates of arbitrary order gradients through differentiation. However, a baseline term that serves as a control variate for reducing variance is currently provided only for first order gradient estimation, limiting the utility of higher-order gradient estimates. To improve the sample efficiency of DiCE, we propose a new base

line term for higher order gradient estimation. This term may be easily included in the objective, and produces unbiased variance-reduced estimators under (automatic) differentiation, without affecting the estimate of the objective itself or of the first order gradient. We provide theoretical analysis and numerical evaluations of our baseline term, which demonstrate that it can dramatically reduce the variance of second order gradient estimators produced by DiCE. This computational tool can be easily used to estimate second order gradients with unprecedented efficiency wherever automatic differentiation is utilised, and has the potential to unlock applications of higher order gradients in reinforcement learning and meta-learning.

\*\*\*\*\*

Probabilistic Neural-Symbolic Models for Interpretable Visual Question Answering  
Ramakrishna Vedantam, Stefan Lee, Marcus Rohrbach, Dhruv Batra, Devi Parikh

We propose a new class of probabilistic neural-symbolic models for visual question answering (VQA) that provide interpretable explanations of their decision making in the form of programs, given a small annotated set of human programs. The key idea of our approach is to learn a rich latent space which effectively propagates program annotations from known questions to novel questions. We do this by formalizing prior work on VQA, called module networks (Andreas, 2016) as discrete, structured, latent variable models on the joint distribution over questions and answers given images, and devise a procedure to train the model effectively.

Our results on a dataset of compositional questions about SHAPES (Andreas, 2016) show that our model generates more interpretable programs and obtains better accuracy on VQA in the low-data regime than prior work.

\*\*\*\*\*

Policy Optimization via Stochastic Recursive Gradient Algorithm

Huizhuo Yuan, Chris Junchi Li, Yuhao Tang, Yuren Zhou

In this paper, we propose the Stochastic Recursive Gradient Policy Optimization (SARAPO) algorithm which is a novel variance reduction method on Trust Region Policy Optimization (TRPO). The algorithm incorporates the Stochastic Recursive Gradient algorithm (SARAH) into the TRPO framework. Compared with the existing Stochastic Variance Reduced Policy Optimization (SVRPO), our algorithm is more stable in the variance. Furthermore, by theoretical analysis the ordinary differential equation and the stochastic differential equation (ODE/SDE) of SARAH, we analyze its convergence property and stability. Our experiments demonstrate its performance on a variety of benchmark tasks. We show that our algorithm gets better improvement in each iteration and matches or even outperforms SVRPO and TRPO.

\*\*\*\*\*

Learning shared manifold representation of images and attributes for generalized zero-shot learning

Masahiro Suzuki, Yusuke Iwasawa, Yutaka Matsuo

Many of the zero-shot learning methods have realized predicting labels of unseen images by learning the relations between images and pre-defined class-attributes. However, recent studies show that, under the more realistic generalized zero-shot learning (GZSL) scenarios, these approaches severely suffer from the issue of biased prediction, i.e., their classifier tends to predict all the examples from both seen and unseen classes as one of the seen classes. The cause of this problem is that they cannot properly learn a mapping to the representation space generalized to the unseen classes since the training set does not include any unseen class information. To solve this, we propose a concept to learn a mapping that embeds both images and attributes to the shared representation space that can be generalized even for unseen classes by interpolating from the information of seen classes, which we refer to as shared manifold learning. Furthermore, we propose modality invariant variational autoencoders, which can perform shared manifold learning by training variational autoencoders with both images and attributes as inputs. The empirical validation of well-known datasets in GZSL shows that our method achieves the significantly superior performances to the existing relation-based studies.

\*\*\*\*\*

## Single Shot Neural Architecture Search Via Direct Sparse Optimization

Xinbang Zhang,Zehao Huang,Naiyan Wang

Recently Neural Architecture Search (NAS) has aroused great interest in both academia and industry, however it remains challenging because of its huge and non-continuous search space. Instead of applying evolutionary algorithm or reinforcement learning as previous works, this paper proposes a Direct Sparse Optimization NAS (DSO-NAS) method. In DSO-NAS, we provide a novel model pruning view to NAS problem. In specific, we start from a completely connected block, and then introduce scaling factors to scale the information flow between operations. Next, we impose sparse regularizations to prune useless connections in the architecture. Lastly, we derive an efficient and theoretically sound optimization method to solve it. Our method enjoys both advantages of differentiability and efficiency, therefore can be directly applied to large datasets like ImageNet. Particularly, On CIFAR-10 dataset, DSO-NAS achieves an average test error 2.84%, while on the ImageNet dataset DSO-NAS achieves 25.4% test error under 600M FLOPs with 8 GPUs in 18 hours.

\*\*\*\*\*

## Continual Learning via Explicit Structure Learning

Xilai Li,Yingbo Zhou,Tianfu Wu,Richard Socher,Caiming Xiong

Despite recent advances in deep learning, neural networks suffer catastrophic forgetting when tasks are learned sequentially. We propose a conceptually simple and general framework for continual learning, where structure optimization is considered explicitly during learning. We implement this idea by separating the structure and parameter learning. During structure learning, the model optimizes for the best structure for the current task. The model learns when to reuse or modify structure from previous tasks, or create new ones when necessary. The model parameters are then estimated with the optimal structure. Empirically, we found that our approach leads to sensible structures when learning multiple tasks continuously. Additionally, catastrophic forgetting is also largely alleviated from explicit learning of structures. Our method also outperforms all other baselines on the permuted MNIST and split CIFAR datasets in continual learning setting.

\*\*\*\*\*

## Neural TTS Stylization with Adversarial and Collaborative Games

Shuang Ma,Daniel McDuff,Yale Song

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 ( $x_{\text{txt}}$ ,  $x_{\text{aud}}$ ) by encouraging its output to reconstruct  $x_{\text{aud}}$ . The synthesized audio waveform is expected to contain the verbal content of  $x_{\text{txt}}$  and the auditory style of  $x_{\text{aud}}$ . 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).

\*\*\*\*\*

## Local Binary Pattern Networks for Character Recognition

Jeng-Hau Lin,Yunfan Yang,Rajesh K. Gupta,Zhuowen Tu

Memory and computation efficient deep learning architectures are crucial to the continued proliferation of machine learning capabilities to new platforms and systems. Binarization of operations in convolutional neural networks has shown promising results in reducing the model size and computing efficiency.

In this paper, we tackle the character recognition problem using a strategy different from the existing literature by proposing local binary pattern networks or LBPNet that can learn and perform bit-wise operations in an end-to-end fashion. LBPNet uses local binary comparisons and random projection in place of conventional convolution (or approximation of convolution) operations, providing important means to improve memory and speed efficiency that is particularly suited for small footprint devices and hardware accelerators. These operations can be implemented efficiently on different platforms including direct hardware implementation. LBPNet demonstrates its particular advantage on the character classification task where the content is composed of strokes. We applied LBPNet to benchmark datasets like MNIST, SVHN, DHCD, ICDAR, and Chars74K and observed encouraging results.

\*\*\*\*\*

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

Yukun Ding, Jinglan Liu, Jinjun Xiong, Yiyu Shi

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  $\mathcal{O}(\log^5(1/\epsilon))$  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.

\*\*\*\*\*

## A CASE STUDY ON OPTIMAL DEEP LEARNING MODEL FOR UAVS

Chandan Kumar, Subrahmanyam Vaddi, Aishwarya Sarkar

Over the passage of time Unmanned Autonomous Vehicles (UAVs), especially Autonomous flying drones grabbed a lot of attention in Artificial Intelligence. Since electronic technology is getting smaller, cheaper and more efficient, huge advancement in the study of UAVs has been observed recently. From monitoring floods, discerning the spread of algae in water bodies to detecting forest trail

, their application is far and wide. Our work is mainly focused on autonomous flying drones where we establish a case study towards efficiency, robustness and accuracy

of UAVs where we showed our results well supported through experiments.

We provide details of the software and hardware architecture used in the study. We

further discuss about our implementation algorithms and present experiments that provide a comparison between three different state-of-the-art algorithms namely TrailNet, InceptionResnet and MobileNet in terms of accuracy, robustness, power consumption and inference time. In our study, we have shown that MobileNet has produced better results with very less computational requirement and power consumption.

We have also reported the challenges we have faced during our work

as well as a brief discussion on our future work to improve safety features and performance.

\*\*\*\*\*

## Poincare Glove: Hyperbolic Word Embeddings

Alexandru Tifrea\*, Gary Becigneul\*, Octavian-Eugen Ganea\*

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-likeness 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

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

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.

\*\*\*\*\*

#### Generative Adversarial Networks for Extreme Learned Image Compression

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, Luc van Gool

We propose a framework for extreme learned image compression based on Generative Adversarial Networks (GANs), obtaining visually pleasing images at significantly lower bitrates than previous methods. This is made possible through our GAN formulation of learned compression combined with a generator/decoder which operates on the full-resolution image and is trained in combination with a multi-scale discriminator. Additionally, if a semantic label map of the original image is available, our method can fully synthesize unimportant regions in the decoded image such as streets and trees from the label map, therefore only requiring the storage of the preserved region and the semantic label map. A user study confirms t



that for low bitrates, our approach is preferred to state-of-the-art methods, even when they use more than double the bits.

\*\*\*\*\*

#### Assessing Generalization in Deep Reinforcement Learning

Charles Packer\*, Katelyn Gao\*, Jernej Kos, Philipp Krahenbuhl, Vladlen Koltun, Dawn Song

Deep reinforcement learning (RL) has achieved breakthrough results on many tasks, but has been shown to be sensitive to system changes at test time. As a result, building deep RL agents that generalize has become an active research area. Our aim is to catalyze and streamline community-wide progress on this problem by providing the first benchmark and a common experimental protocol for investigating generalization in RL. Our benchmark contains a diverse set of environments and our evaluation methodology covers both in-distribution and out-of-distribution generalization. To provide a set of baselines for future research, we conduct a systematic evaluation of state-of-the-art algorithms, including those that specifically tackle the problem of generalization. The experimental results indicate that in-distribution generalization may be within the capacity of current algorithms, while out-of-distribution generalization is an exciting challenge for future work.

\*\*\*\*\*

#### Neural Message Passing for Multi-Label Classification

Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi

Multi-label classification (MLC) is the task of assigning a set of target labels for a given sample. Modeling the combinatorial label interactions in MLC has been a long-haul challenge. Recurrent neural network (RNN) based encoder-decoder models have shown state-of-the-art performance for solving MLC. However, the sequential nature of modeling label dependencies through an RNN limits its ability in parallel computation, predicting dense labels, and providing interpretable results. In this paper, we propose Message Passing Encoder-Decoder (MPED) Networks, aiming to provide fast, accurate, and interpretable MLC. MPED networks model the joint prediction of labels by replacing all RNNs in the encoder-decoder architecture with message passing mechanisms and dispense with autoregressive inference entirely. The proposed models are simple, fast, accurate, interpretable, and structure-agnostic (can be used on known or unknown structured data). Experiments on seven real-world MLC datasets show the proposed models outperform autoregressive RNN models across five different metrics with a significant speedup during training and testing time.

\*\*\*\*\*

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

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, Andrew Gordon Wilson

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

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

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.

\*\*\*\*\*

Correction Networks: Meta-Learning for Zero-Shot Learning

R. Lily Hu, Caiming Xiong, Richard Socher

We propose a model that learns to perform zero-shot classification using a meta-learner that is trained to produce a correction to the output of a previously trained learner. The model consists of two modules: a task module that supplies an initial prediction, and a correction module that updates the initial prediction. The task module is the learner and the correction module is the meta-learner. The correction module is trained in an episodic approach whereby many different task modules are trained on various subsets of the total training data, with the rest being used as unseen data for the correction module. The correction module takes as input a representation of the task module's training data so that the predicted correction is a function of the task module's training data. The correction module is trained to update the task module's prediction to be closer to the target value. This approach leads to state-of-the-art performance for zero-shot classification on natural language class descriptions on the CUB and NAB datasets.

\*\*\*\*\*

Tinkering with black boxes: counterfactuals uncover modularity in generative models

Michel Besserve, Remy Sun, Bernhard Schoelkopf

Deep generative models such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) are important tools to capture and investigate the properties of complex empirical data. However, the complexity of their inner elements makes their functionment challenging to assess and modify. In this respect, these architectures behave as black box models. In order to better understand the function of such networks, we analyze their modularity based on the counterfactual manipulation of their internal variables. Our experiments on the

generation of human faces with VAEs and GANs support that modularity between activation maps distributed over channels of generator architectures is achieved to some degree, can be used to better understand how these systems operate and allow meaningful transformations of the generated images without further training.

erate and edit the content of generated images.

\*\*\*\*\*

Exploiting Cross-Lingual Subword Similarities in Low-Resource Document Classification

Mozhi Zhang, Yoshinari Fujinuma, Jordan Boyd-Graber

Text classification must sometimes be applied in situations with no training data in a target language. However, training data may be available in a related language. We introduce a cross-lingual document classification framework CACO between related language pairs. To best use limited training data, our transfer learning scheme exploits cross-lingual subword similarity by jointly training a character-based embedder and a word-based classifier. The embedder derives vector representations for input words from their written forms, and the classifier makes predictions based on the word vectors. We use a joint character representation for both the source language and the target language, which allows the embedder to generalize knowledge about source language words to target language words with similar forms. We propose a multi-task objective that can further improve the model if additional cross-lingual or monolingual resources are available. CACO models trained under low-resource settings rival cross-lingual word embedding models trained under high-resource settings on related language pairs.

\*\*\*\*\*

DyRep: Learning Representations over Dynamic Graphs

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, Hongyuan Zha

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

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 set up 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

a 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

This paper introduces a new framework for open-domain question answering in which the retriever and the reader *iteratively interact* with each other. The framework is agnostic to the architecture of the machine reading model provided it has *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 *state* of the reader and the *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*. Code and pretrained models are available at <https://github.com/rajarshd/Multi-Step-Reasoning>.

\*\*\*\*\*

Empirically Characterizing Overparameterization Impact on Convergence

Newsha Ardalani, Joel Hestness, Gregory Diamos

A long-held conventional wisdom states that larger models train more slowly when using gradient descent. This work challenges this widely-held belief, showing that larger models can potentially train faster despite the increasing computational requirements of each training step. In particular, we study the effect of network structure (depth and width) on halting time and show that larger models---wider models in particular---take fewer training steps to converge.

We design simple experiments to quantitatively characterize the effect of overparameterization on weight space traversal. Results show that halting time improves when growing model's width for three different applications, and the improvement comes from each factor: The distance from initialized weights to converged weights shrinks with a power-law-like relationship, the average step size grows with a power-law-like relationship, and gradient vectors become more aligned with each other during traversal.

\*\*\*\*\*

Differential Equation Networks

MohamadAli Torkamani, Phillip Wallis

Most deep neural networks use simple, fixed activation functions, such as sigmoids or rectified linear units, regardless of domain or network structure. We introduce differential equation networks, an improvement to modern neural networks in which each neuron learns the particular nonlinear activation function that it requires. We show that enabling each neuron with the ability to learn its own activation function results in a more compact network capable of achieving comparable, if not superior performance when compared to much larger networks. We

also showcase the capability of a differential equation neuron to learn behaviors, such as oscillation, currently only obtainable by a large group of neurons. The ability of differential equation networks to essentially compress a large neural network, without loss of overall performance makes them suitable for on-device applications, where predictions must

be computed locally. Our experimental evaluation of real-world and toy datasets show that differential equation networks outperform fixed activation networks in several areas.

\*\*\*\*\*

Unsupervised Multi-Target Domain Adaptation: An Information Theoretic Approach  
Behnam Gholami, Pritish Sahu, Ognjen (Oggi) Rudovic, Konstantinos Bousmalis, Vladimir Pavlovic

Unsupervised domain adaptation (uDA) models focus on pairwise adaptation settings where there is a single, labeled, source and a single target domain. However, in many real-world settings one seeks to adapt to multiple, but somewhat similar, target domains. Applying pairwise adaptation approaches to this setting may be suboptimal, as they would fail to leverage shared information among the multiple domains. In this work we propose an information theoretic approach for domain adaptation in the novel context of multiple target domains with unlabeled instances and one source domain with labeled instances. Our model aims to find a shared latent space common to all domains, while simultaneously accounting for the remaining private, domain-specific factors. Disentanglement of shared and private information is accomplished using a unified information-theoretic approach, which also serves to provide a stronger link between the latent representations and the observed data. The resulting single model, accompanied by an efficient optimization algorithm, allows simultaneous adaptation from a single source to multiple target domains.

We test our approach on three publicly-available datasets, showing that it outperforms several popular domain adaptation methods.

\*\*\*\*\*

Explicit Recall for Efficient Exploration

Honghua Dong, Jiayuan Mao, Xinyue Cui, Lihong Li

In this paper, we advocate the use of explicit memory for efficient exploration in reinforcement learning. This memory records structured trajectories that have led to interesting states in the past, and can be used by the agent to revisit those states more effectively. In high-dimensional decision making problems, where deep reinforcement learning is considered crucial, our approach provides a simple, transparent and effective way that can be naturally combined with complex, deep learning models. We show how such explicit memory may be used to enhance existing exploration algorithms such as intrinsically motivated ones and count-based ones, and demonstrate our method's advantages in various simulated environments.

\*\*\*\*\*

Improving machine classification using human uncertainty measurements

Ruairidh M. Battleday, Joshua C. Peterson, Thomas L. Griffiths

As deep CNN classifier performance using ground-truth labels has begun to asymptote at near-perfect levels, a key aim for the field is to extend training paradigms to capture further useful structure in natural image data and improve model robustness and generalization. In this paper, we present a novel natural image benchmark for making this extension, which we call CIFAR10H. This new dataset comprises a human-derived, full distribution over labels for each image of the CIFAR10 test set, offering the ability to assess the generalization of state-of-the-art CIFAR10 models, as well as investigate the effects of including this information in model training. We show that classification models trained on CIFAR10 do not generalize as well to our dataset as it does to traditional extensions, and that models fine-tuned using our label information are able to generalize better to related datasets, complement popular data augmentation schemes, and provide robustness to adversarial attacks. We explain these improvements in terms of better empirical approximations to the expected loss function over natural images and their categories in the visual world.

\*\*\*\*\*

Capsule Graph Neural Network

Zhang Xinyi, Lihui Chen

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

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

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

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.

\*\*\*\*\*

#### Graph Generation via Scattering

Dongmian Zou, Gilad Lerman

Generative networks have made it possible to generate meaningful signals such as images and texts from simple noise. Recently, generative methods based on GAN and VAE were developed for graphs and graph signals. However, the mathematical properties of these methods are unclear, and training good generative models is difficult. This work proposes a graph generation model that uses a recent adaptation of Mallat's scattering transform to graphs. The proposed model is naturally composed of an encoder and a decoder. The encoder is a Gaussianized graph scattering transform, which is robust to signal and graph manipulation. The decoder is a simple fully connected network that is adapted to specific tasks, such as link prediction, signal generation on graphs and full graph and signal generation. The training of our proposed system is efficient since it is only applied to the decoder and the hardware requirement is moderate. Numerical results demonstrate state-of-the-art performance of the proposed system for both link prediction and graph and signal generation. These results are in contrast to experience with Euclidean data, where it is difficult to form a generative scattering network that performs as well as state-of-the-art methods. We believe that this is because of the discrete and simpler nature of graph applications, unlike the more complex and high-frequency nature of Euclidean data, in particular, of some natural images.

\*\*\*\*\*

#### Learning Physics Priors for Deep Reinforcement Learning

Yilun Du, Karthik Narasimhan

While model-based deep reinforcement learning (RL) holds great promise for sample efficiency and generalization, learning an accurate dynamics model is challenging and often requires substantial interactions with the environment. Further, a wide variety of domains have dynamics that share common foundations like the laws of physics, which are rarely exploited by these algorithms. Humans often acquire such physics priors that allow us to easily adapt to the dynamics of any environment. In this work, we propose an approach to learn such physics priors and incorporate them into an RL agent. Our method involves pre-training a frame predictor on raw videos and then using it to initialize the dynamics prediction model on a target task. Our prediction model, SpatialNet, is designed to implicitly capture localized physical phenomena and interactions. We show the value of incorporating this prior through empirical experiments on two different domains - a newly created PhysWorld and games from the Atari benchmark, outperforming competitive approaches and demonstrating effective transfer learning.

\*\*\*\*\*

#### Composing Complex Skills by Learning Transition Policies

Youngwoon Lee\*, Shao-Hua Sun\*, Sriram Somasundaram, Edward S. Hu, Joseph J. Lim

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 reward

s 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

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.

\*\*\*\*\*

Manifold Alignment via Feature Correspondence

Jay S. Stanley III, Guy Wolf, Smita Krishnaswamy

We propose a novel framework for combining datasets via alignment of their associated intrinsic dimensions. Our approach assumes that the two datasets are sampled from a common latent space, i.e., they measure equivalent systems. Thus, we expect there to exist a natural (albeit unknown) alignment of the data manifolds associated with the intrinsic geometry of these datasets, which are perturbed by measurement artifacts in the sampling process. Importantly, we do not assume any individual correspondence (partial or complete) between data points. Instead, we rely on our assumption that a subset of data features have correspondence across datasets. We leverage this assumption to estimate relations between intrinsic manifold dimensions, which are given by diffusion map coordinates over each of the datasets. We compute a correlation matrix between diffusion coordinates of the datasets by considering graph (or manifold) Fourier coefficients of corresponding data features. We then orthogonalize this correlation matrix to form an isometric transformation between the diffusion maps of the datasets. Finally, we apply this transformation to the diffusion coordinates and construct a unified diffusion geometry of the datasets together. We show that this approach successfully corrects misalignment artifacts, and allows for integrated data.

\*\*\*\*\*

ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware

Han Cai, Ligeng Zhu, Song Han

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<sup>4</sup> 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 (grows 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



et. 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

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

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 Dinulescu, Douglas Eck

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.

\*\*\*\*\*

SynonymNet: Multi-context Bilateral Matching for Entity Synonyms

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, Philip S. Yu

Being able to automatically discover synonymous entities from a large free-text corpus has transformative effects on structured knowledge discovery. Existing works either require structured annotations, or fail to incorporate context information effectively, which lower the efficiency of information usage. In this paper, we propose a framework for synonym discovery from free-text corpus without structured annotation. As one of the key components in synonym discovery, we introduce a novel neural network model SynonymNet to determine whether or not two given entities are synonym with each other. Instead of using entities features, SynonymNet makes use of multiple pieces of contexts in which the entity is mentioned, and compares the context-level similarity via a bilateral matching schema to determine synonymity. Experimental results demonstrate that the proposed model achieves state-of-the-art results on both generic and domain-specific synonym datasets: Wiki+Freebase, PubMed+UMLS and MedBook+MKG, with up to 4.16% improvement in terms of Area Under the Curve (AUC) and 3.19% in terms of Mean Average Precision (MAP) compare to the best baseline method.

\*\*\*\*\*

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

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.

\*\*\*\*\*

Learning to Progressively Plan

Xinyun Chen, Yuandong Tian

For problem solving, making reactive decisions based on problem description is fast but inaccurate, while search-based planning using heuristics gives better solutions but could be exponentially slow. In this paper, we propose a new approach that improves an existing solution by iteratively picking and rewriting its local components until convergence. The rewriting policy employs a neural network trained with reinforcement learning. We evaluate our approach in two domains: job scheduling and expression simplification. Compared to common effective heuristics, baseline deep models and search algorithms, our approach efficiently gives solutions with higher quality.

\*\*\*\*\*

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

Mikael Henaff, Alfredo Canziani, Yann LeCun

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

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

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.

\*\*\*\*\*

## The Natural Language Decathlon: Multitask Learning as Question Answering

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher

Deep learning has improved performance on many natural language processing (NLP) tasks individually.

However, general NLP models cannot emerge within a paradigm that focuses on the particularities of a single metric, dataset, and task.

We introduce the Natural Language Decathlon (decaNLP), a challenge that spans ten tasks:

question answering, machine translation, summarization, natural language inference, sentiment analysis, semantic role labeling, relation extraction, goal-oriented dialogue, semantic parsing, and commonsense pronoun resolution.

We cast all tasks as question answering over a context.

Furthermore, we present a new multitask question answering network (MQAN) that jointly learns all tasks in decaNLP without any task-specific modules or parameters more effectively than sequence-to-sequence and reading comprehension baselines.

MQAN shows improvements in transfer learning for machine translation and named entity recognition, domain adaptation for sentiment analysis and natural language inference, and zero-shot capabilities for text classification.

We demonstrate that the MQAN's multi-pointer-generator decoder is key to this success and that performance further improves with an anti-curriculum training strategy.

Though designed for decaNLP, MQAN also achieves state of the art results on the WikiSQL semantic parsing task in the single-task setting.

We also release code for procuring and processing data, training and evaluating models, and reproducing all experiments for decaNLP.

\*\*\*\*\*

Contingency-Aware Exploration in Reinforcement Learning

Jongwook Choi, Yijie Guo, Marcin Moczulski, Junhyuk Oh, Neal Wu, Mohammad Norouzi, Honglak Lee

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.

\*\*\*\*\*

Using GANs for Generation of Realistic City-Scale Ride Sharing/Hailing Data Sets  
Abhinav Jauhri, Brad Stocks, Jian Hui Li, Koichi Yamada, John Paul Shen

This paper focuses on the synthetic generation of human mobility data in urban areas. We present a novel and scalable application of Generative Adversarial Networks (GANs) for modeling and generating human mobility data. We leverage actual ride requests from ride sharing/hailing services from four major cities in the US to train our GANs model. Our model captures the spatial and temporal variability of the ride-request patterns observed for all four cities on any typical day and over any typical week. Previous works have succinctly characterized the spatial and temporal properties of human mobility data sets using the fractal dimensionality and the densification power law, respectively, which we utilize to validate our GANs-generated synthetic data sets. Such synthetic data sets can avoid privacy concerns and be extremely useful for researchers and policy makers on urban mobility and intelligent transportation.

\*\*\*\*\*

## Neural Graph Evolution: Towards Efficient Automatic Robot Design

Tingwu Wang, Yuhao Zhou, Sanja Fidler, Jimmy Ba

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

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.

\*\*\*\*\*

## Ain't Nobody Got Time for Coding: Structure-Aware Program Synthesis from Natural Language

Jakub Bednarek, Karol Piaskowski, Krzysztof Krawiec

Program synthesis from natural language (NL) is practical for humans and, once technically feasible, would significantly facilitate software development and revolutionize end-user programming. We present SAPS, an end-to-end neural network capable of mapping relatively complex, multi-sentence NL specifications to snippets of executable code. The proposed architecture relies exclusively on neural components, and is built upon a tree2tree autoencoder trained on abstract syntax trees, combined with a pretrained word embedding and a bi-directional multi-layer LSTM for NL processing. The decoder features a doubly-recurrent LSTM with a novel signal propagation scheme and soft attention mechanism. When applied to a large dataset of problems proposed in a previous study, SAPS performs on par with o

r better than the method proposed there, producing correct programs in over 90% of cases. In contrast to other methods, it does not involve any non-neural components to post-process the resulting programs, and uses a fixed-dimensional latent representation as the only link between the NL analyzer and source code generator.

\*\*\*\*\*

#### Provable Defenses against Spatially Transformed Adversarial Inputs: Impossibility and Possibility Results

Xinyang Zhang, Yifan Huang, Chanh Nguyen, Shouling Ji, Ting Wang

One intriguing property of neural networks is their inherent vulnerability to adversarial inputs, which are maliciously crafted samples to trigger target networks to misbehave. The state-of-the-art attacks generate adversarial inputs using either pixel perturbation or spatial transformation. Thus far, several provable defenses have been proposed against pixel perturbation-based attacks; yet, little is known about whether such solutions exist for spatial transformation-based attacks. This paper bridges this striking gap by conducting the first systematic study on provable defenses against spatially transformed adversarial inputs. Our findings convey mixed messages. On the impossibility side, we show that such defenses may not exist in practice: for any given networks, it is possible to find legitimate inputs and imperceptible transformations to generate adversarial inputs that force arbitrarily large errors. On the possibility side, we show that it is still feasible to construct adversarial training methods to significantly improve the resilience of networks against adversarial inputs over empirical datasets. We believe our findings provide insights for designing more effective defenses against spatially transformed adversarial inputs.

\*\*\*\*\*

#### Simple Black-box Adversarial Attacks

Chuan Guo, Jacob R. Gardner, Yurong You, Andrew G. Wilson, Kilian Q. Weinberger

The construction of adversarial images is a search problem in high dimensions within a small region around a target image. The goal is to find an imperceptibly modified image that is misclassified by a target model. In the black-box setting, only sporadic feedback is provided through occasional model evaluations. In this paper we provide a new algorithm whose search strategy is based on an intriguingly simple iterative principle: We randomly pick a low frequency component of the discrete cosine transform (DCT) and either add or subtract it to the target image. Model evaluations are only required to identify whether an operation decreases the adversarial loss. Despite its simplicity, the proposed method can be used for targeted and untargeted attacks --- resulting in previously unprecedented query efficiency in both settings. We require a median of 600 black-box model queries (ResNet-50) to produce an adversarial ImageNet image, and we successfully attack Google Cloud Vision with 2500 median queries, averaging to a cost of only \$3 per image. We argue that our proposed algorithm should serve as a strong baseline for future adversarial black-box attacks, in particular because it is extremely fast and can be implemented in less than 20 lines of PyTorch code.

\*\*\*\*\*

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

Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, Antonio Torralba

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

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 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

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 to control self-assembling morphologies: a study of generalization via modularity

Deepak Pathak, Chris Lu, Trevor Darrell, Philip Isola, Alexei A. Efros

Much of contemporary sensorimotor learning assumes that one is already given a complex agent (e.g., a robotic arm) and the goal is to learn to control it. In contrast, this paper investigates a modular co-evolution strategy: a collection of primitive agents learns to self-assemble into increasingly complex collectives in order to solve control tasks. Each primitive agent consists of a limb and a neural controller. Limbs may choose to link up to form collectives, with linking being treated as a dynamic action. When two limbs link, a joint is added between them, actuated by the 'parent' limb's controller. This forms a new 'single' agent, which may further link with other agents. In this way, complex morphologies can emerge, controlled by a policy whose architecture is in explicit correspondence with the morphology. In experiments, we demonstrate that agents with these modular and dynamic topologies generalize better to test-time environments compared to static and monolithic baselines. Project videos are available at <https://doubleblindICLR19.github.io/self-assembly/>

\*\*\*\*\*

Learning Procedural Abstractions and Evaluating Discrete Latent Temporal Structure

Karan Goel, Emma Brunskill

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.

\*\*\*\*\*

Stochastic Quantized Activation: To prevent Overfitting in Fast Adversarial Training

Wonjun Yoon, Jisuk Park, Daeshik Kim

Existing neural networks are vulnerable to "adversarial examples"---created by adding maliciously designed small perturbations in inputs to induce a misclassification by the networks. The most investigated defense strategy is adversarial training which augments training data with adversarial examples. However, applying single-step adversaries in adversarial training does not support the robustness of the networks, instead, they will even make the networks to be overfitted. In contrast to the single-step, multi-step training results in the state-of-the-art performance on MNIST and CIFAR10, yet it needs a massive amount of time. Therefore, we propose a method, Stochastic Quantized Activation (SQA) that solves overfitting problems in single-step adversarial training and fastly achieves the robustness comparable to the multi-step. SQA attenuates the adversarial effects by providing random selectivity to activation functions and allows the network to learn robustness with only single-step training. Throughout the experiment, our method demonstrates the state-of-the-art robustness against one of the strongest white-box attacks as PGD training, but with much less computational cost. Finally, we visualize the learning process of the network with SQA to handle strong adversaries, which is different from existing methods.

\*\*\*\*\*

Learning from Noisy Demonstration Sets via Meta-Learned Suitability Assessor

Te-Lin Wu, Jaedong Hwang, Jingyun Yang, Shaofan Lai, Carl Vondrick, Joseph J. Lim

A noisy and diverse demonstration set may hinder the performances of an agent aiming to acquire certain skills via imitation learning. However, state-of-the-art imitation learning algorithms often assume the optimality of the given demonstration set.

In this paper, we address such optimal assumption by learning only from the most suitable demonstrations in a given set. Suitability of a demonstration is estimated by whether imitating it produce desirable outcomes for achieving the goals of the tasks. For more efficient demonstration suitability assessments, the learning agent should be capable of imitating a demonstration as quick as possible, which shares similar spirit with fast adaptation in the meta-learning regime. Our framework, thus built on top of Model-Agnostic Meta-Learning, evaluates how desirable the imitated outcomes are, after adaptation to each demonstration in the set. The resulting assessments hence enable us to select suitable demonstration subsets for acquiring better imitated skills. The videos related to our experiments are available at: <https://sites.google.com/view/deepdj>

\*\*\*\*\*

Object-Oriented Model Learning through Multi-Level Abstraction

Guangxiang Zhu, Jianhao Wang, ZhiZhou Ren, Chongjie Zhang

Object-based approaches for learning action-conditioned dynamics has demonstrated promise for generalization and interpretability. However, existing approaches suffer from structural limitations and optimization difficulties for common environments with multiple dynamic objects. In this paper, we present a novel self-supervised learning framework, called Multi-level Abstraction Object-oriented Predictor (MAOP), for learning object-based dynamics models from raw visual observ



ations. MAOP employs a three-level learning architecture that enables efficient dynamics learning for complex environments with a dynamic background. We also design a spatial-temporal relational reasoning mechanism to support instance-level dynamics learning and handle partial observability. Empirical results show that MAOP significantly outperforms previous methods in terms of sample efficiency and generalization over novel environments that have multiple controllable and uncontrollable dynamic objects and different static object layouts. In addition, MAOP learns semantically and visually interpretable disentangled representations.

\*\*\*\*\*

#### Learning to Design RNA

Frederic Runge, Danny Stoll, Stefan Falkner, Frank Hutter

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.

\*\*\*\*\*

#### Accelerated Sparse Recovery Under Structured Measurements

Ke Li, Jitendra Malik

Extensive work on compressed sensing has yielded a rich collection of sparse recovery algorithms, each making different tradeoffs between recovery condition and computational efficiency. In this paper, we propose a unified framework for accelerating various existing sparse recovery algorithms without sacrificing recovery guarantees by exploiting structure in the measurement matrix. Unlike fast algorithms that are specific to particular choices of measurement matrices where the columns are Fourier or wavelet filters for example, the proposed approach works on a broad range of measurement matrices that satisfy a particular property. We precisely characterize this property, which quantifies how easy it is to accelerate sparse recovery for the measurement matrix in question. We also derive the time complexity of the accelerated algorithm, which is sublinear in the signal length in each iteration. Moreover, we present experimental results on real world data that demonstrate the effectiveness of the proposed approach in practice.

\*\*\*\*\*

#### REVISITING NEGATIVE TRANSFER USING ADVERSARIAL LEARNING

Saneem Ahmed Chemmengath, Samarth Bharadwaj, Suranjana Samanta, Karthik Sankaranarayanan

An unintended consequence of feature sharing is the model fitting to correlated tasks within the dataset, termed negative transfer. In this paper, we revisit the problem of negative transfer in multitask setting and find that its corrosive effects are applicable to a wide range of linear and non-linear models, including neural networks. We first study the effects of negative transfer in a principled way and show that previously proposed counter-measures are insufficient, particularly for trainable features. We propose an adversarial training approach to mitigate the effects of negative transfer by viewing the problem in a domain adaptation setting. Finally, empirical results on attribute prediction multi-task

on AWA and CUB datasets further validate the need for correcting negative sharing in an end-to-end manner.

\*\*\*\*\*

#### Mean Replacement Pruning

Utku Evci, Nicolas Le Roux, Pablo Castro, Leon Bottou

Pruning units in a deep network can help speed up inference and training as well as reduce the size of the model. We show that bias propagation is a pruning technique which consistently outperforms the common approach of merely removing units, regardless of the architecture and the dataset. We also show how a simple adaptation to an existing scoring function allows us to select the best units to prune. Finally, we show that the units selected by the best performing scoring functions are somewhat consistent over the course of training, implying the dead parts of the network appear during the stages of training.

\*\*\*\*\*

#### P<sup>2</sup>IR: Universal Deep Node Representation via Partial Permutation Invariant Set Functions

Shupeng Gui, Xiangliang Zhang, Shuang Qiu, Mingrui Wu, Jieping Ye, Ji Liu

Graph node representation learning is a central problem in social network analysis, aiming to learn the vector representation for each node in a graph. The key problem is how to model the dependence of each node to its neighbor nodes since the neighborhood can uniquely characterize a graph. Most existing approaches rely on defining the specific neighborhood dependence as the computation mechanism of representations, which may exclude important subtle structures within the graph and dependence among neighbors. Instead, we propose a novel graph node embedding method (namely P<sup>2</sup>IR) via developing a novel notion, namely partial permutation invariant set function} to learn those subtle structures. Our method can 1) learn an arbitrary form of the representation function from the neighborhood, without losing any potential dependence structures, 2) automatically decide the significance of neighbors at different distances, and 3) be applicable to both homogeneous and heterogeneous graph embedding, which may contain multiple types of nodes. Theoretical guarantee for the representation capability of our method has been proved for general homogeneous and heterogeneous graphs. Evaluation results on benchmark data sets show that the proposed P<sup>2</sup>IR outperforms the state-of-the-art approaches on producing node vectors for classification tasks.

\*\*\*\*\*

#### Accelerating first order optimization algorithms

Angelos Katrakis, Roger Nankambou

There exist several stochastic optimization algorithms. However in most cases, it is difficult to tell for a particular problem which will be the best optimizer to choose as each of them are good. Thus, we present a simple and intuitive technique, when applied to first order optimization algorithms, is able to improve the speed of convergence and reaches a better minimum for the loss function compared to the original algorithms. The proposed solution modifies the update rule, based on the variation of the direction of the gradient during training. We conducted several tests with Adam and AMSGrad on two different datasets. The preliminary results show that the proposed technique improves the performance of existing optimization algorithms and works well in practice.

\*\*\*\*\*

#### Accelerated Gradient Flow for Probability Distributions

Amirhossein Taghvaei, Prashant G. Mehta

This paper presents a methodology and numerical algorithms for constructing accelerated gradient flows on the space of probability distributions. In particular, we extend the recent variational formulation of accelerated gradient methods in Wibisono2016 from vector valued variables to probability distributions. The variational problem is modeled as a mean-field optimal control problem. The maximum principle of optimal control theory is used to derive Hamilton's equations for the optimal gradient flow. The Hamilton's equation are shown to achieve the accelerated form of density transport from any initial probability distribution to a target probability distribution. A quantitative estimate on the asymptotic convergence rate is provided based on a Lyapunov function construction, when the

objective functional is displacement convex. Two numerical approximations are presented to implement the Hamilton's equations as a system of  $N$  interacting particles. The continuous limit of the Nesterov's algorithm is shown to be a special case with  $N=1$ . The algorithm is illustrated with numerical examples.

\*\*\*\*\*

#### Cost-Sensitive Robustness against Adversarial Examples

Xiao Zhang, David Evans

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

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.

\*\*\*\*\*

#### Modulating transfer between tasks in gradient-based meta-learning

Erin Grant, Ghassen Jerfel, Katherine Heller, Thomas L. Griffiths

Learning-to-learn or meta-learning leverages data-driven inductive bias to increase the efficiency of learning on a novel task. This approach encounters difficulty when transfer is not mutually beneficial, for instance, when tasks are sufficiently dissimilar or change over time. Here, we use the connection between gradient-based meta-learning and hierarchical Bayes to propose a mixture of hierarchical Bayesian models over the parameters of an arbitrary function approximator such as a neural network. Generalizing the model-agnostic meta-learning (MAML) algorithm, we present a stochastic expectation maximization procedure to jointly estimate parameter initializations for gradient descent as well as a latent assignment of tasks to initializations. This approach better captures the diversity of training tasks as opposed to consolidating inductive biases into a single set of hyperparameters. Our experiments demonstrate better generalization on the standard miniImageNet benchmark for 1-shot classification. We further derive a novel and scalable non-parametric variant of our method that captures the evolution of a task distribution over time as demonstrated on a set of few-shot regression tasks.

\*\*\*\*\*

#### Outlier Detection from Image Data

Lei Cao, Yizhou Yan, Samuel Madden, Elke Rundensteiner

Modern applications from Autonomous Vehicles to Video Surveillance generate massive amounts of image data. In this work we propose a novel image outlier detection approach (IOD for short) that leverages the cutting-edge image classifier to discover outliers without using any labeled outlier. We observe that although intuitively the confidence that a convolutional neural network (CNN) has that an image belongs to a particular class could serve as outlieriness measure to each image, directly applying this confidence to detect outlier does not work well. This is because CNN often has high confidence on an outlier image that does not belong to any target class due to its generalization ability that ensures the high accuracy in classification. To solve this issue, we propose a Deep Neural Forest-based approach that harmonizes the contradictory requirements of accurately classifying images and correctly detecting the outlier images. Our experiments using several benchmark image datasets including MNIST, CIFAR-10, CIFAR-100, and SVHN demonstrate the effectiveness of our IOD approach for outlier detection, capturing more than 90% of outliers generated by injecting one image dataset into another, while still preserving the classification accuracy of the multi-class classification problem.

\*\*\*\*\*

A Variational Inequality Perspective on Generative Adversarial Networks

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, Simon Lacoste-Julien

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.

\*\*\*\*\*

Measuring Density and Similarity of Task Relevant Information in Neural Representations

Danish Pruthi, Mansi Gupta, Nitish Kumar Kulkarni, Graham Neubig, Eduard Hovy

Neural models achieve state-of-the-art performance due to their ability to extract salient features useful to downstream tasks. However, our understanding of how this task-relevant information is included in these networks is still incomplete. In this paper, we examine two questions (1) how densely is information included in extracted representations, and (2) how similar is the encoding of relevant information between related tasks. We propose metrics to measure information density and cross-task similarity, and perform an extensive analysis in the domain of natural language processing, using four varieties of sentence representation and 13 tasks. We also demonstrate how the proposed analysis tools can find immediate use in choosing tasks for transfer learning.

\*\*\*\*\*

Multiple-Attribute Text Rewriting

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, Y-Lan Boureau

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 t

he 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.

\*\*\*\*\*

#### Experience replay for continual learning

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, Greg Wayne

Continual learning is the problem of learning new tasks or knowledge while protecting old knowledge and ideally generalizing from old experience to learn new tasks faster. Neural networks trained by stochastic gradient descent often degrade on old tasks when trained successively on new tasks with different data distributions. This phenomenon, referred to as catastrophic forgetting, is considered a major hurdle to learning with non-stationary data or sequences of new tasks, and prevents networks from continually accumulating knowledge and skills. We examine this issue in the context of reinforcement learning, in a setting where an agent is exposed to tasks in a sequence. Unlike most other work, we do not provide an explicit indication to the model of task boundaries, which is the most general circumstance for a learning agent exposed to continuous experience. While various methods to counteract catastrophic forgetting have recently been proposed, we explore a straightforward, general, and seemingly overlooked solution - that of using experience replay buffers for all past events - with a mixture of on- and off-policy learning, leveraging behavioral cloning. We show that this strategy can still learn new tasks quickly yet can substantially reduce catastrophic forgetting in both Atari and DMLab domains, even matching the performance of methods that require task identities. When buffer storage is constrained, we confirm that a simple mechanism for randomly discarding data allows a limited size buffer to perform almost as well as an unbounded one.

\*\*\*\*\*

#### Search-Guided, Lightly-supervised Training of Structured Prediction Energy Networks

Amirmohammad Rooshenas, Dongxu Zhang, Gopal Sharma, Andrew McCallum

In structured output prediction tasks, labeling ground-truth training output is often expensive. However, for many tasks, even when the true output is unknown, we can evaluate predictions using a scalar reward function, which may be easily assembled from human knowledge or non-differentiable pipelines. But searching through the entire output space to find the best output with respect to this reward function is typically intractable. In this paper, we instead use efficient truncated randomized search in this reward function to train structured prediction energy networks (SPENs), which provide efficient test-time inference using gradient-based search on a smooth, learned representation of the score landscape, and have previously yielded state-of-the-art results in structured prediction. In particular, this truncated randomized search in the reward function yields previously unknown local improvements, providing effective supervision to SPENs, avoiding their traditional need for labeled training data.

\*\*\*\*\*

#### Multi-Agent Dual Learning

Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, Tie-Yan Liu

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.

\*\*\*\*\*

## Doubly Sparse: Sparse Mixture of Sparse Experts for Efficient Softmax Inference

Shun Liao, Ting Chen, Tian Lin, Chong Wang, Dengyong Zhou

Computations for the softmax function in neural network models are expensive when the number of output classes is large. This can become a significant issue in both training and inference for such models. In this paper, we present Doubly Sparse Softmax (DS-Softmax), Sparse Mixture of Sparse of Sparse Experts, to improve the efficiency for softmax inference. During training, our method learns a two-level class hierarchy by dividing entire output class space into several partially overlapping experts. Each expert is responsible for a learned subset of the output class space and each output class only belongs to a small number of those experts. During inference, our method quickly locates the most probable expert to compute small-scale softmax. Our method is learning-based and requires no knowledge of the output class partition space a priori. We empirically evaluate our method on several real-world tasks and demonstrate that we can achieve significant computation reductions without loss of performance.

\*\*\*\*\*

## Optimistic Acceleration for Optimization

Jun-Kun Wang, Xiaoyun Li, Ping Li

We consider new variants of optimization algorithms. Our algorithms are based on the observation that mini-batch of stochastic gradients in consecutive iterations do not change drastically and consequently may be predictable. Inspired by the similar setting in online learning literature called Optimistic Online learning, we propose two new optimistic algorithms for AMSGrad and Adam, respectively, by exploiting the predictability of gradients. The new algorithms combine the idea of momentum method, adaptive gradient method, and algorithms in Optimistic Online learning, which leads to speed up in training deep neural nets in practice.

\*\*\*\*\*

## The Variational Deficiency Bottleneck

Pradeep Kr. Banerjee, Guido Montufar

We introduce a bottleneck method for learning data representations based on channel deficiency, rather than the more traditional information sufficiency. A variational upper bound allows us to implement this method efficiently. The bound itself is bounded above by the variational information bottleneck objective, and the two methods coincide in the regime of single-shot Monte Carlo approximations. The notion of deficiency provides a principled way of approximating complicated channels by relatively simpler ones. The deficiency of one channel w.r.t. another has an operational interpretation in terms of the optimal risk gap of decision problems, capturing classification as a special case. Unsupervised generalizations are possible, such as the deficiency autoencoder, which can also be formulated in a variational form. Experiments demonstrate that the deficiency bottleneck can provide advantages in terms of minimal sufficiency as measured by information bottleneck curves, while retaining a good test performance in classification and reconstruction tasks.

\*\*\*\*\*

## AntMan: Sparse Low-Rank Compression To Accelerate RNN Inference

Samyam Rajbhandari, Harsh Shrivastava, Yuxiong He

Wide adoption of complex RNN based models is hindered by their inference performance, cost and memory requirements. To address this issue, we develop AntMan, combining structured sparsity with low-rank decomposition synergistically, to reduce model computation, size and execution time of RNNs while attaining desired accuracy. AntMan extends knowledge distillation based training to learn the compressed models efficiently. Our evaluation shows that AntMan offers up to 100x computation reduction with less than 1pt accuracy drop for language and machine reading comprehension models. Our evaluation also shows that for a given accuracy target, AntMan produces 5x smaller models than the state-of-art. Lastly, we show that AntMan offers super-linear speed gains compared to theoretical speedup, demonstrating its practical value on commodity hardware.

\*\*\*\*\*

## Learning sparse relational transition models

Victoria Xia,Zi Wang,Kelsey Allen,Tom Silver,Leslie Pack Kaelbling

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.

\*\*\*\*\*

## IB-GAN: Disentangled Representation Learning with Information Bottleneck GAN

Insu Jeon,Wonkwang Lee,Gunhee Kim

We present a novel architecture of GAN for a disentangled representation learning. The new model architecture is inspired by Information Bottleneck (IB) theory thereby named IB-GAN. IB-GAN objective is similar to that of InfoGAN but has a crucial difference; a capacity regularization for mutual information is adopted, thanks to which the generator of IB-GAN can harness a latent representation in a disentangled and interpretable manner. To facilitate the optimization of IB-GAN in practice, a new variational upper-bound is derived. With experiments on CelebA, 3DChairs, and dSprites datasets, we demonstrate that the visual quality of samples generated by IB-GAN is often better than those by  $\beta$ -VAEs. Moreover, IB-GAN achieves much higher disentanglement metrics score than  $\beta$ -VAEs or InfoGAN on the dSprites dataset.

\*\*\*\*\*

## SPIGAN: Privileged Adversarial Learning from Simulation

Kuan-Hui Lee,German Ros,Jie Li,Adrien Gaidon

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.

\*\*\*\*\*

## MahiNet: A Neural Network for Many-Class Few-Shot Learning with Class Hierarchy

Lu Liu,Tianyi Zhou,Guodong Long,Jing Jiang,Chengqi Zhang

We study many-class few-shot (MCFS) problem in both supervised learning and meta-learning scenarios. Compared to the well-studied many-class many-shot and few-class few-shot problems, MCFS problem commonly occurs in practical applications but is rarely studied. MCFS brings new challenges because it needs to distinguish between many classes, but only a few samples per class are available for training. In this paper, we propose 'memory-augmented hierarchical-classification network (MahiNet)' for MCFS learning. It addresses the 'many-class' problem by exploring the class hierarchy, e.g., the coarse-class label that covers a subset of fine classes, which helps to narrow down the candidates for the fine class and is cheaper to obtain. MahiNet uses a convolutional neural network (CNN) to extract features, and integrates a memory-augmented attention module with a multi-layer perceptron (MLP) to produce the probabilities over coarse and fine classes. While the MLP extends the linear classifier, the attention module extends a KNN classifier, both together targeting the 'few-shot' problem. We design different training strategies of MahiNet for supervised learning and meta-learning. Moreover, we propose two novel benchmark datasets 'mcfsImageNet' (as a subset of ImageNet) and 'mcfsOmniplot' (re-splitting Omniplot) specifically for MCFS pro

blem. In experiments, we show that MahiNet outperforms several state-of-the-art models on MCFS classification tasks in both supervised learning and meta-learning scenarios.

\*\*\*\*\*

#### Interpretable Continual Learning

Tameem Adel, Cuong V. Nguyen, Richard E. Turner, Zoubin Ghahramani, Adrian Weller

We present a framework for interpretable continual learning (ICL). We show that explanations of previously performed tasks can be used to improve performance on future tasks. ICL generates a good explanation of a finished task, then uses this to focus attention on what is important when facing a new task. The ICL idea is general and may be applied to many continual learning approaches. Here we focus on the variational continual learning framework to take advantage of its flexibility and efficacy in overcoming catastrophic forgetting. We use saliency maps to provide explanations of performed tasks and propose a new metric to assess their quality. Experiments show that ICL achieves state-of-the-art results in terms of overall continual learning performance as measured by average classification accuracy, and also in terms of its explanations, which are assessed qualitatively and quantitatively using the proposed metric.

\*\*\*\*\*

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

Zaiyi Chen, Zhuoning Yuan, Jinfeng Yi, Bowen Zhou, Enhong Chen, Tianbao Yang

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  $\{\text{ada}\}$  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

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 t



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

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.

\*\*\*\*\*

#### Adaptive Mixture of Low-Rank Factorizations for Compact Neural Modeling

Ting Chen, Ji Lin, Tian Lin, Song Han, Chong Wang, Denny Zhou

Modern deep neural networks have a large amount of weights, which make them difficult to deploy on computation constrained devices such as mobile phones. One common approach to reduce the model size and computational cost is to use low-rank factorization to approximate a weight matrix. However, performing standard low-rank factorization with a small rank can hurt the model expressiveness and significantly decrease the performance. In this work, we propose to use a mixture of multiple low-rank factorizations to model a large weight matrix, and the mixture coefficients are computed dynamically depending on its input. We demonstrate the effectiveness of the proposed approach on both language modeling and image classification tasks. Experiments show that our method not only improves the computation efficiency but also maintains (sometimes outperforms) its accuracy compared with the full-rank counterparts.

\*\*\*\*\*

#### Functional Bayesian Neural Networks for Model Uncertainty Quantification

Nanyang Ye, Zhanxing Zhu

In this paper, we extend the Bayesian neural network to functional Bayesian neural network with functional Monte Carlo methods that use the samples of functionals instead of samples of networks' parameters for inference to overcome the curse of dimensionality for uncertainty quantification. Based on the previous work on Riemannian Langevin dynamics, we propose the stochastic gradient functional Riemannian dynamics for training functional Bayesian neural network. We show the effectiveness and efficiency of our proposed approach with various experiments.

\*\*\*\*\*

#### Fake Sentence Detection as a Training Task for Sentence Encoding

Viresh Ranjan, Heeyoung Kwon, Niranjan Balasubramanian, Minh Hoai

Sentence encoders are typically trained on generative language modeling tasks with large unlabeled datasets. While these encoders achieve strong results on many sentence-level tasks, they are difficult to train with long training cycles.

.

We introduce fake sentence detection as a new discriminative training task for learning sentence encoders. We automatically generate fake sentences by corrupting original sentences from a source collection and train the encoders to produce representations that are effective at detecting fake sentences. This binary classification task turns to be quite efficient for training sentence encoders. We compare a basic BiLSTM encoder trained on this task with strong sentence encoding models (Skipthought and FastSent) trained on a language modeling task. We find that the BiLSTM trains much faster on fake sentence detection (20 hours inst

ead of weeks) using smaller amounts of data (1M instead of 64M sentences). Further analysis shows the learned representations also capture many syntactic and semantic properties expected from good sentence representations.

\*\*\*\*\*

#### Multi-Grained Entity Proposal Network for Named Entity Recognition

Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, Philip S. Yu

In this paper, we focus on a new Named Entity Recognition (NER) task, i.e., the Multi-grained NER task. This task aims to simultaneously detect both fine-grained and coarse-grained entities in sentences. Correspondingly, we develop a novel Multi-grained Entity Proposal Network (MGEPN). Different from traditional NER models which regard NER as a sequential labeling task, MGEPN provides a new method that proposes entity candidates in the Proposal Network and classifies entities into different categories in the Classification Network. All possible entity candidates including fine-grained ones and coarse-grained ones are proposed in the Proposal Network, which enables the MGEPN model to identify multi-grained entities. In order to better identify named entities and determine their categories, context information is utilized and transferred from the Proposal Network to the Classification Network during the learning process. A novel Entity-Context attention mechanism is also introduced to help the model focus on entity-related context information. Experiments show that our model can obtain state-of-the-art performance on two real-world datasets for both the Multi-grained NER task and the traditional NER task.

\*\*\*\*\*

#### A Model Cortical Network for Spatiotemporal Sequence Learning and Prediction

Jielin Qiu, Ge Huang, Tai Sing Lee

In this paper we developed a hierarchical network model, called Hierarchical Prediction Network (HPNet) to understand how spatiotemporal memories might be learned and encoded in a representational hierarchy for predicting future video frames. The model is inspired by the feedforward, feedback and lateral recurrent circuits in the mammalian hierarchical visual system. It assumes that spatiotemporal memories are encoded in the recurrent connections within each level and between different levels of the hierarchy. The model contains a feed-forward path that computes and encodes spatiotemporal features of successive complexity and a feedback path that projects interpretation from a higher level to the level below. Within each level, the feed-forward path and the feedback path intersect in a recurrent gated circuit that integrates their signals as well as the circuit's internal memory states to generate a prediction of the incoming signals. The network learns by comparing the incoming signals with its prediction, updating its internal model of the world by minimizing the prediction errors at each level of the hierarchy in the style of {\em predictive self-supervised learning}. The network processes data in blocks of video frames rather than a frame-to-frame basis. This allows it to learn relationships among movement patterns, yielding state-of-the-art performance in long range video sequence predictions in benchmark datasets. We observed that hierarchical interaction in the network introduces sensitivity to memories of global movement patterns even in the population representation of the units in the earliest level. Finally, we provided neurophysiological evidence, showing that neurons in the early visual cortex of awake monkeys exhibit very similar sensitivity and behaviors. These findings suggest that predictive self-supervised learning might be an important principle for representational learning in the visual cortex.

\*\*\*\*\*

#### ISA-VAE: Independent Subspace Analysis with Variational Autoencoders

Jan Stühmer, Richard Turner, Sebastian Nowozin

Recent work has shown increased interest in using the Variational Autoencoder (VAE) framework to discover interpretable representations of data in an unsupervised way. These methods have focussed largely on modifying the variational cost function to achieve this goal. However, we show that methods like beta-VAE simplify the tendency of variational inference to underfit causing pathological over-pruning and over-orthogonalization of learned components. In this paper we take a

complementary approach: to modify the probabilistic model to encourage structured latent variable representations to be discovered. Specifically, the standard VAE probabilistic model is unidentifiable: the likelihood of the parameters is invariant under rotations of the latent space. This means there is no pressure to identify each true factor of variation with a latent variable.

We therefore employ a rich prior distribution, akin to the ICA model, that breaks the rotational symmetry.

Extensive quantitative and qualitative experiments demonstrate that the proposed prior mitigates the trade-off introduced by modified cost functions like beta-VAE and TCVAE between reconstruction loss and disentanglement. The proposed prior allows to improve these approaches with respect to both disentanglement and reconstruction quality significantly over the state of the art.

\*\*\*\*\*

On the Trajectory of Stochastic Gradient Descent in the Information Plane

Emilio Rafael Balda, Arash Behboodi, Rudolf Mathar

Studying the evolution of information theoretic quantities during Stochastic Gradient Descent (SGD) learning of Artificial Neural Networks (ANNs) has gained popularity in recent years.

Nevertheless, these type of experiments require estimating mutual information and entropy which becomes intractable for moderately large problems. In this work we propose a framework for understanding SGD learning in the information plane which consists of observing entropy and conditional entropy of the output labels of ANN. Through experimental results and theoretical justifications it is shown that, under some assumptions, the SGD learning trajectories appear to be similar for different ANN architectures. First, the SGD learning is modeled as a Hidden Markov Process (HMP) whose entropy tends to increase to the maximum. Then, it is shown that the SGD learning trajectory appears to move close to the shortest path between the initial and final joint distributions in the space of probability measures equipped with the total variation metric. Furthermore, it is shown that the trajectory of learning in the information plane can provide an alternative for observing the learning process, with potentially richer information about the learning than the trajectories in training and test error.

\*\*\*\*\*

NOODL: Provable Online Dictionary Learning and Sparse Coding

Sirisha Rambhatla, Xingguo Li, Jarvis Haupt

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.

\*\*\*\*\*

Neural Model-Based Reinforcement Learning for Recommendation

Xinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Le Song

There are great interests as well as many challenges in applying reinforcement learning (RL) to recommendation systems. In this setting, an online user is the environment; neither the reward function nor the environment dynamics are clearly

defined, making the application of RL challenging.

In this paper, we propose a novel model-based reinforcement learning framework for recommendation systems, where we develop a generative adversarial network to imitate user behavior dynamics and learn her reward function. Using this user model as the simulation environment, we develop a novel DQN algorithm to obtain a combinatorial recommendation policy which can handle a large number of candidate items efficiently. In our experiments with real data, we show this generative adversarial user model can better explain user behavior than alternatives, and the RL policy based on this model can lead to a better long-term reward for the user and higher click rate for the system.

\*\*\*\*\*

RelGAN: Relational Generative Adversarial Networks for Text Generation

Weili Nie, Nina Narodytska, Ankit Patel

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

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.

\*\*\*\*\*

Characterizing Attacks on Deep Reinforcement Learning

Chaowei Xiao, Xinlei Pan, Warren He, Bo Li, Jian Peng, Mingjie Sun, Jinfeng Yi, Mingyan Liu, Dawn Song.

Deep Reinforcement learning (DRL) has achieved great success in various applications, such as playing computer games and controlling robotic manipulation. However

er, recent studies show that machine learning models are vulnerable to adversarial examples, which are carefully crafted instances that aim to mislead learning models to make arbitrarily incorrect prediction, and raised severe security concerns. DRL has been attacked by adding perturbation to each observed frame. However, such observation based attacks are not quite realistic considering that it would be hard for adversaries to directly manipulate pixel values in practice. Therefore, we propose to understand the vulnerabilities of DRL from various perspectives and provide a throughout taxonomy of adversarial perturbation against DRL, and we conduct the first experiments on unexplored parts of this taxonomy. In addition to current observation based attacks against DRL, we propose attacks based on the actions and environment dynamics. Among these experiments, we introduce a novel sequence-based attack to attack a sequence of frames for real-time scenarios such as autonomous driving, and the first targeted attack that perturbs environment dynamics to let the agent fail in a specific way. We show empirically that our sequence-based attack can generate effective perturbations in a blackbox setting in real time with a small number of queries, independent of episode length. We conduct extensive experiments to compare the effectiveness of different attacks with several baseline attack methods in several game playing, robotics control, and autonomous driving environments.

\*\*\*\*\*

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

Pramod Kaushik Mudrakarta, Mark Sandler, Andrey Zhmoginov, Andrew Howard

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

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

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.

ning (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.

\*\*\*\*\*

ACTRCE: Augmenting Experience via Teacher’s Advice

Yuhuai Wu, Harris Chan, Jamie Kiros, Sanja Fidler, Jimmy Ba

Sparse reward is one of the most challenging problems in reinforcement learning (RL). Hindsight Experience Replay (HER) attempts to address this issue by converting a failure experience to a successful one by relabeling the goals. Despite its effectiveness, HER has limited applicability because it lacks a compact and universal goal representation. We present Augmenting experience via Teacher’s advice (ACTRCE), an efficient reinforcement learning technique that extends the HER framework using natural language as the goal representation. We first analyze the differences among goal representation, and show that ACTRCE can efficiently solve difficult reinforcement learning problems in challenging 3D navigation tasks, whereas HER with non-language goal representation failed to learn. We also show that with language goal representations, the agent can generalize to unseen instructions, and even generalize to instructions with unseen lexicons. We further demonstrate it is crucial to use hindsight advice to solve challenging tasks, but we also found that little amount of hindsight advice is sufficient for the learning to take off, showing the practical aspect of the method.

\*\*\*\*\*

Learning to Augment Influential Data

Donghoon Lee, Chang D. Yoo

Data augmentation is a technique to reduce overfitting and to improve generalization by increasing the number of labeled data samples by performing label preserving transformations; however, it is currently conducted in a trial and error manner. A composition of predefined transformations, such as rotation, scaling and cropping, is performed on training samples, and its effect on performance over test samples can only be empirically evaluated and cannot be predicted. This paper considers an influence function which predicts how generalization is affected by a particular augmented training sample in terms of validation loss. The influence function provides an approximation of the change in validation loss without comparing the performance which includes and excludes the sample in the training process. A differentiable augmentation model that generalizes the conventional composition of predefined transformations is also proposed. The differentiable augmentation model and reformulation of the influence function allow the parameters of the augmented model to be directly updated by backpropagation to minimize the validation loss. The experimental results show that the proposed method provides better generalization over conventional data augmentation methods.

\*\*\*\*\*

Unsupervised Exploration with Deep Model-Based Reinforcement Learning

Kurtland Chua, Rowan McAllister, Roberto Calandra, Sergey Levine

Reinforcement learning (RL) often requires large numbers of trials to solve a single specific task. This is in sharp contrast to human and animal learning: humans and animals can use past experience to acquire an understanding about the world, which they can then use to perform new tasks with minimal additional learning. In this work, we study how an unsupervised exploration phase can be used to build up such prior knowledge, which can then be utilized in a second phase to perform new tasks, either directly without any additional exploration, or through minimal fine-tuning. A critical question with this approach is: what kind of knowledge should be transferred from the unsupervised phase to the goal-directed phase? We argue that model-based RL offers an appealing solution. By transferring models, which are task-agnostic, we can perform new tasks without any additional learning at all. However, this relies on having a suitable exploration method during unsupervised training, and a model-based RL method that can effectively utilize modern high-capacity parametric function classes, such as deep neural networks.

orks. We show that both challenges can be addressed by representing model-uncertainty, which can both guide exploration in the unsupervised phase and ensure that the errors in the model are not exploited by the planner in the goal-directed phase. We illustrate, on simple simulated benchmark tasks, that our method can perform various goal-directed skills on the first attempt, and can improve further with fine-tuning, exceeding the performance of alternative exploration methods.

\*\*\*\*\*

#### INTERPRETABLE CONVOLUTIONAL FILTER PRUNING

Zhuwei Qin, Fuxun Yu, Chenchen Liu, Xiang Chen

The sophisticated structure of Convolutional Neural Network (CNN) allows for outstanding performance, but at the cost of intensive computation. As significant

redundancies inevitably present in such a structure, many works have been proposed

to prune the convolutional filters for computation cost reduction. Although extremely effective, most works are based only on quantitative characteristics of

the convolutional filters, and highly overlook the qualitative interpretation of individual

filter's specific functionality. In this work, we interpreted the functionality and redundancy of the convolutional filters from different perspectives, and proposed

a functionality-oriented filter pruning method. With extensive experimental results, we proved the convolutional filters' qualitative significance regardless of

magnitude, demonstrated significant neural network redundancy due to repetitive filter functions, and analyzed the filter functionality defection under inappropriate

retraining process. Such an interpretable pruning approach not only offers outstanding

computation cost optimization over previous filter pruning methods, but also interprets filter pruning process.

\*\*\*\*\*

#### 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

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

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.

\*\*\*\*\*

Combining adaptive algorithms and hypergradient method: a performance and robustness study

Akram Erragabi, Nicolas Le Roux

Wilson et al. (2017) showed that, when the stepsize schedule is properly designed, stochastic gradient generalizes better than ADAM (Kingma & Ba, 2014). In light of recent work on hypergradient methods (Baydin et al., 2018), we revisit these claims to see if such methods close the gap between the most popular optimizers. As a byproduct, we analyze the true benefit of these hypergradient methods compared to more classical schedules, such as the fixed decay of Wilson et al. (2017). In particular, we observe they are of marginal help since their performance varies significantly when tuning their hyperparameters. Finally, as robustness is a critical quality of an optimizer, we provide a sensitivity analysis of these gradient based optimizers to assess how challenging their tuning is.

\*\*\*\*\*

Novel positional encodings to enable tree-structured transformers

Vighnesh Leonardo Shiv, Chris Quirk

With interest in program synthesis and similarly favored problems rapidly increasing, neural models optimized for tree-domain problems are of great value. In the sequence domain, transformers can learn relationships across arbitrary pairs of positions with less bias than recurrent models. Under the intuition that a similar property would be beneficial in the tree domain, we propose a method to extend transformers to tree-structured inputs and/or outputs. Our approach abstracts transformer's default sinusoidal positional encodings, allowing us to substitute in a novel custom positional encoding scheme that represents node positions within a tree. We evaluated our model in tree-to-tree program translation and sequence-to-tree semantic parsing settings, achieving superior performance over the vanilla transformer model on several tasks.

\*\*\*\*\*

Adversarial Audio Super-Resolution with Unsupervised Feature Losses

Sung Kim, Visvesh Sathe

Neural network-based methods have recently demonstrated state-of-the-art results on image synthesis and super-resolution tasks, in particular by using variants of generative adversarial networks (GANs) with supervised feature losses. Nevertheless, previous feature loss formulations rely on the availability of large auxiliary classifier networks, and labeled datasets that enable such classifiers to be trained. Furthermore, there has been comparatively little work to explore the applicability of GAN-based methods to domains other than images and video. In this work we explore a GAN-based method for audio processing, and develop a convolutional neural network architecture to perform audio super-resolution. In addi



tion to several new architectural building blocks for audio processing, a key component of our approach is the use of an autoencoder-based loss that enables training in the GAN framework, with feature losses derived from unlabeled data. We explore the impact of our architectural choices, and demonstrate significant improvements over previous works in terms of both objective and perceptual quality.

\*\*\*\*\*

#### Visual Imitation with a Minimal Adversary

Scott Reed, Yusuf Aytar, Ziyu Wang, Tom Paine, Aäron van den Oord, Tobias Pfaff, Sergio Gomez, Alexander Novikov, David Budden, Oriol Vinyals

High-dimensional sparse reward tasks present major challenges for reinforcement learning agents. In this work we use imitation learning to address two of these challenges: how to learn a useful representation of the world e.g. from pixels, and how to explore efficiently given the rarity of a reward signal? We show that adversarial imitation can work well even in this high dimensional observation space. Surprisingly the adversary itself, acting as the learned reward function, can be tiny, comprising as few as 128 parameters, and can be easily trained using the most basic GAN formulation. Our approach removes limitations present in most contemporary imitation approaches: requiring no demonstrator actions (only video), no special initial conditions or warm starts, and no explicit tracking of any single demo. The proposed agent can solve a challenging robot manipulation task of block stacking from only video demonstrations and sparse reward, in which the non-imitating agents fail to learn completely. Furthermore, our agent learns much faster than competing approaches that depend on hand-crafted, staged dense reward functions, and also better compared to standard GAIL baselines. Finally, we develop a new adversarial goal recognizer that in some cases allows the agent to learn stacking without any task reward, purely from imitation.

\*\*\*\*\*

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

Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, Ruitong Huang

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

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

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 Fréchet 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.

\*\*\*\*\*

#### Making Convolutional Networks Shift-Invariant Again

Richard Zhang

Modern convolutional networks are not shift-invariant, despite their convolutional nature: small shifts in the input can cause drastic changes in the internal feature maps and output. In this paper, we isolate the cause -- the downsampling operation in convolutional and pooling layers -- and apply the appropriate signal processing fix -- low-pass filtering before downsampling. This simple architectural modification boosts the shift-equivariance of the internal representations and consequently, shift-invariance of the output. Importantly, this is achieved while maintaining downstream classification performance. In addition, incorporating the inductive bias of shift-invariance largely removes the need for shift-biased data augmentation. Lastly, we observe that the modification induces spatially-smoother learned convolutional kernels. Our results suggest that this classical signal processing technique has a place in modern deep networks.

\*\*\*\*\*

#### Characterizing Audio Adversarial Examples Using Temporal Dependency

Zhuolin Yang, Bo Li, Pin-Yu Chen, Dawn Song

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 discriminative 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.

\*\*\*\*\*

#### Text Infilling

Wanrong Zhu, Zhiting Hu, Eric P. Xing

Recent years have seen remarkable progress of text generation in different contexts

xts, including the most common setting of generating text from scratch, the increasingly popular paradigm of retrieval and editing, and others. Text infilling, which fills missing text portions of a sentence or paragraph, is also of numerous use in real life. Previous work has focused on restricted settings, by either assuming single word per missing portion, or limiting to single missing portion to the end of text. This paper studies the general task of text infilling, where the input text can have an arbitrary number of portions to be filled, each of which may require an arbitrary unknown number of tokens.

We develop a self-attention model with segment-aware position encoding for precise global context modeling.

We further create a variety of supervised data by masking out text in different domains with varying missing ratios and mask strategies. Extensive experiments show the proposed model performs significantly better than other methods, and generates meaningful text patches.

\*\*\*\*\*

GANSynth: Adversarial Neural Audio Synthesis

Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, Adam Roberts

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

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  $O(1/\sqrt{T})$ .

\*\*\*\*\*

Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, Pieter Abbeel

Flow-based generative models are powerful exact likelihood models with efficient sampling and inference.

Despite their computational efficiency, flow-based models generally have much worse density modeling performance compared to state-of-the-art autoregressive models. In this paper, we investigate and improve upon three limiting design choices employed by flow-based models in prior work: the use of uniform noise for dequantization, the use of inexpressive affine flows, and the use of purely convolutional conditioning networks in coupling layers. Based on our findings, we propose

e Flow++, a new flow-based model that is now the state-of-the-art non-autoregressive model for unconditional density estimation on standard image benchmarks. Our work has begun to close the significant performance gap that has so far existed between autoregressive models and flow-based models.

\*\*\*\*\*

Beyond Greedy Ranking: Slate Optimization via List-CVAE

Ray Jiang, Sven Gowal, Yuqiu Qian, Timothy Mann, Danilo J. Rezende

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 interdependencies. 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.

\*\*\*\*\*

Intrinsic Social Motivation via Causal Influence in Multi-Agent RL

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, Nando de Freitas

We derive a new intrinsic social motivation for multi-agent reinforcement learning (MARL), in which agents are rewarded for having causal influence over another agent's actions, where causal influence is assessed using counterfactual reasoning. The reward does not depend on observing another agent's reward function, and is thus a more realistic approach to MARL than taken in previous work. We show that the causal influence reward is related to maximizing the mutual information between agents' actions. We test the approach in challenging social dilemma environments, where it consistently leads to enhanced cooperation between agents and higher collective reward. Moreover, we find that rewarding influence can lead agents to develop emergent communication protocols. Therefore, we also employ influence to train agents to use an explicit communication channel, and find that it leads to more effective communication and higher collective reward. Finally, we show that influence can be computed by equipping each agent with an internal model that predicts the actions of other agents. This allows the social influence reward to be computed without the use of a centralised controller, and as such represents a significantly more general and scalable inductive bias for MARL with independent agents.

\*\*\*\*\*

Model Compression with Generative Adversarial Networks

Ruishan Liu, Nicolo Fusi, Lester Mackey

More accurate machine learning models often demand more computation and memory at test time, making them difficult to deploy on CPU- or memory-constrained devices. Model compression (also known as distillation) alleviates this burden by training a less expensive student model to mimic the expensive teacher model while maintaining most of the original accuracy. However, when fresh data is unavailable for the compression task, the teacher's training data is typically reused, leading to suboptimal compression. In this work, we propose to augment the compression dataset with synthetic data from a generative adversarial network (GAN) designed to approximate the training data distribution. Our GAN-assisted model compression (GAN-MC) significantly improves student accuracy for expensive models such as deep neural networks and large random forests on both image and tabular data

tasets. Building on these results, we propose a comprehensive metric—the Compression Score—to evaluate the quality of synthetic datasets based on their induced model compression performance. The Compression Score captures both data diversity and discriminability, and we illustrate its benefits over the popular Inception Score in the context of image classification.

\*\*\*\*\*

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

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.

\*\*\*\*\*

Architecture Compression

Anubhav Ashok

In this paper we propose a novel approach to model compression termed Architecture Compression. Instead of operating on the weight or filter space of the network like classical model compression methods, our approach operates on the architecture space. A 1-D CNN encoder/decoder is trained to learn a mapping from discrete architecture space to a continuous embedding and back. Additionally, this embedding is jointly trained to regress accuracy and parameter count in order to incorporate information about the architecture's effectiveness on the dataset. During the compression phase, we first encode the network and then perform gradient descent in continuous space to optimize a compression objective function that maximizes accuracy and minimizes parameter count. The final continuous feature is then mapped to a discrete architecture using the decoder. We demonstrate the merits of this approach on visual recognition tasks such as CIFAR-10/100, FMNIST and SVHN and achieve a greater than 20x compression on CIFAR-10.

\*\*\*\*\*

ATTENTIVE EXPLAINABILITY FOR PATIENT TEMPORAL EMBEDDING

Daby Sow, Mohamed Ghalwash, Zach Shahn, Sanjoy Dey, Moulay Draidia, Li-wei Lehmann  
Learning explainable patient temporal embeddings from observational data has mostly ignored the use of RNN architecture that excel in capturing temporal data dependencies but at the expense of explainability. This paper addresses this problem by introducing and applying an information theoretic approach to estimate the degree of explainability of such architectures. Using a communication paradigm, we formalize metrics of explainability by estimating the amount of information that an AI model needs to convey to a human end user to explain and rationalize its outputs. A key aspect of this work is to model human prior knowledge at the receiving end and measure the lack of explainability as a deviation from human prior knowledge. We apply this paradigm to medical concept representation problems by regularizing loss functions of temporal autoencoders according to the deriv

ed explainability metrics to guide the learning process towards models producing explainable outputs. We illustrate the approach with convincing experimental results for the generation of explainable temporal embeddings for critical care patient data.

\*\*\*\*\*

#### Context Dependent Modulation of Activation Function

Long Sha, Jonathan Schwarcz, Pengyu Hong

We propose a modification to traditional Artificial Neural Networks (ANNs), which provides the ANNs with new aptitudes motivated by biological neurons. Biological neurons work far beyond linearly summing up synaptic inputs and then transforming the integrated information. A biological neuron changes firing modes according to peripheral factors (e.g., neuromodulators) as well as intrinsic ones.

Our modification connects a new type of ANN nodes, which mimic the function of biological neuromodulators and are termed modulators, to enable other traditional ANN nodes to adjust their activation sensitivities in run-time based on their input patterns. In this manner, we enable the slope of the activation function to be context dependent. This modification produces statistically significant improvements in comparison with traditional ANN nodes in the context of Convolutional Neural Networks and Long Short-Term Memory networks.

\*\*\*\*\*

#### Transfer Value or Policy? A Value-centric Framework Towards Transferrable Continuous Reinforcement Learning

Xingchao Liu, Tongzhou Mu, Hao Su

Transferring learned knowledge from one environment to another is an important step towards practical reinforcement learning (RL). In this paper, we investigate the problem of transfer learning across environments with different dynamics while accomplishing the same task in the continuous control domain. We start by illustrating the limitations of policy-centric methods (policy gradient, actor-critic, etc.) when transferring knowledge across environments. We then propose a general model-based value-centric (MVC) framework for continuous RL. MVC learns a dynamics approximator and a value approximator simultaneously in the source domain, and makes decision based on both of them. We evaluate MVC against popular baselines on 5 benchmark control tasks in a training from scratch setting and a transfer learning setting. Our experiments demonstrate MVC achieves comparable performance with the baselines when it is trained from scratch, while it significantly surpasses them when it is used in the transfer setting.

\*\*\*\*\*

#### Uncovering Surprising Behaviors in Reinforcement Learning via Worst-case Analysis

Avraham Ruderman, Richard Everett, Bristy Sikder, Hubert Soyer, Jonathan Uesato, Ananya Kumar, Charlie Beattie, Pushmeet Kohli

Reinforcement learning agents are typically trained and evaluated according to their performance averaged over some distribution of environment settings. But does the distribution over environment settings contain important biases, and do these lead to agents that fail in certain cases despite high average-case performance? In this work, we consider worst-case analysis of agents over environment settings in order to detect whether there are directions in which agents may have failed to generalize. Specifically, we consider a 3D first-person task where agents must navigate procedurally generated mazes, and where reinforcement learning agents have recently achieved human-level average-case performance. By optimizing over the structure of mazes, we find that agents can suffer from catastrophic failures, failing to find the goal even on surprisingly simple mazes, despite their impressive average-case performance. Additionally, we find that these failures transfer between different agents and even significantly different architectures. We believe our findings highlight an important role for worst-case analysis in identifying whether there are directions in which agents have failed to generalize. Our hope is that the ability to automatically identify failures of generalization will facilitate development of more general and robust agents. To this end, we report initial results on enriching training with settings causing fa

ilure.

\*\*\*\*\*

#### Spherical CNNs on Unstructured Grids

Chiyu Max Jiang, Jingwei Huang, Karthik Kashinath, Prabhat, Philip Marcus, Matthias Nießner

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

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

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

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 Per turb-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.

\*\*\*\*\*

Small steps and giant leaps: Minimal Newton solvers for Deep Learning

Joao Henriques, Sebastien Ehrhardt, Samuel Albanie, Andrea Vedaldi

We propose a fast second-order method that can be used as a drop-in replacement for current deep learning solvers. Compared to stochastic gradient descent (SGD), it only requires two additional forward-mode automatic differentiation operations per iteration, which has a computational cost comparable to two standard forward passes and is easy to implement. Our method addresses long-standing issues with current second-order solvers, which invert an approximate Hessian matrix every iteration exactly or by conjugate-gradient methods, procedures that are much slower than a SGD step. Instead, we propose to keep a single estimate of the gradient projected by the inverse Hessian matrix, and update it once per iteration with just two passes over the network. This estimate has the same size and is similar to the momentum variable that is commonly used in SGD. No estimate of the Hessian is maintained.

We first validate our method, called CurveBall, on small problems with known solutions (noisy Rosenbrock function and degenerate 2-layer linear networks), where current deep learning solvers struggle. We then train several large models on CIFAR and ImageNet, including ResNet and VGG-f networks, where we demonstrate faster convergence with no hyperparameter tuning. We also show our optimiser's generality by testing on a large set of randomly-generated architectures.

\*\*\*\*\*

DEEP-TRIM: REVISITING L1 REGULARIZATION FOR CONNECTION PRUNING OF DEEP NETWORK

Chih-Kuan Yeh, Ian E.H. Yen, Hong-You Chen, Chun-Pei Yang, Shou-De Lin, Pradeep Ravikumar

State-of-the-art deep neural networks (DNNs) typically have tens of millions of parameters, which might not fit into the upper levels of the memory hierarchy, thus increasing the inference time and energy consumption significantly, and prohibiting their use on edge devices such as mobile phones. The compression of DNN models has therefore become an active area of research recently, with connection pruning emerging as one of the most successful strategies. A very natural approach is to prune connections of DNNs via  $\ell_1$  regularization, but recent empirical investigations have suggested that this does not work as well in the context of DNN compression. In this work, we revisit this simple strategy and analyze it rigorously, to show that: (a) any stationary point of an  $\ell_1$ -regularized layerwise-pruning objective has its number of non-zero elements bounded by the number of penalized prediction logits, regardless of the strength of the regularization; (b) successful pruning highly relies on an accurate optimization solver, and there is a trade-off between compression speed and distortion of prediction accuracy, controlled by the strength of regularization. Our theoretical results thus suggest that  $\ell_1$  pruning could be successful provided we use an accurate optimization solver. We corroborate this in our experiments, where we show that simple  $\ell_1$  regularization with an Adamax-L1(cumulative) solver gives pruning ratio competitive to the state-of-the-art.

\*\*\*\*\*

Neural Rendering Model: Joint Generation and Prediction for Semi-Supervised Learning

Nhat Ho, Tan Nguyen, Ankit B. Patel, Anima Anandkumar, Michael I. Jordan, Richard G. Baraniuk

Unsupervised and semi-supervised learning are important problems that are especially challenging with complex data like natural images. Progress on these problems would accelerate if we had access to appropriate generative models under which to pose the associated inference tasks. Inspired by the success of Convolutional Neural Networks (CNNs) for supervised prediction in images, we design the Neural Rendering Model (NRM), a new hierarchical probabilistic generative model who



se inference calculations correspond to those in a CNN. The NRM introduces a small set of latent variables at each level of the model and enforces dependencies among all the latent variables via a conjugate prior distribution. The conjugate prior yields a new regularizer for learning based on the paths rendered in the generative model for training CNNs—the Rendering Path Normalization (RPN). We demonstrate that this regularizer improves generalization both in theory and in practice. Likelihood estimation in the NRM yields the new Max-Min cross entropy training loss, which suggests a new deep network architecture—the Max-Min network—which exceeds or matches the state-of-art for semi-supervised and supervised learning on SVHN, CIFAR10, and CIFAR100.

\*\*\*\*\*

NICE: noise injection and clamping estimation for neural network quantization  
Chaim Baskin, Natan Liss, Yoav Chai, Evgenii Zheltonozhskii, Eli Schwartz, Raja Girayes, Avi Mendelson, Alexander M. Bronstein

Convolutional Neural Networks (CNN) are very popular in many fields including computer vision, speech recognition, natural language processing, to name a few. Though deep learning leads to groundbreaking performance in these domains, the networks used are very demanding computationally and are far from real-time even on a GPU, which is not power efficient and therefore does not suit low power systems such as mobile devices. To overcome this challenge, some solutions have been proposed for quantizing the weights and activations of these networks, which accelerate the runtime significantly. Yet, this acceleration comes at the cost of a larger error. The NICE method proposed in this work trains quantized neural networks by noise injection and a learned clamping, which improve the accuracy. This leads to state-of-the-art results on various regression and classification tasks, e.g., ImageNet classification with architectures such as ResNet-18/34/50 with low as 3-bit weights and 3-bit activations. We implement the proposed solution on an FPGA to demonstrate its applicability for low power real-time applications.

\*\*\*\*\*

Trajectory VAE for multi-modal imitation

Xiaoyu Lu, Jan Stuehmer, Katja Hofmann

We address the problem of imitating multi-modal expert demonstrations in sequential decision making problems. In many practical applications, for example video games, behavioural demonstrations are readily available that contain multi-modal structure not captured by typical existing imitation learning approaches. For example, differences in the observed players' behaviours may be representative of different underlying playstyles.

In this paper, we use a generative model to capture different emergent playstyles in an unsupervised manner, enabling the imitation of a diverse range of distinct behaviours. We utilise a variational autoencoder to learn an embedding of the different types of expert demonstrations on the trajectory level, and jointly learn a latent representation with a policy. In experiments on a range of 2D continuous control problems representative of Minecraft environments, we empirically demonstrate that our model can capture a multi-modal structured latent space from the demonstrated behavioural trajectories.

\*\*\*\*\*

Approximation capability of neural networks on sets of probability measures and tree-structured data

Tomáš Pevný, Vojtěch Kovařík

This paper extends the proof of density of neural networks in the space of continuous (or even measurable) functions on Euclidean spaces to functions on compact sets of probability measures.

By doing so the work parallels a more than a decade old results on mean-map embedding of probability measures in reproducing kernel Hilbert spaces.

The work has wide practical consequences for multi-instance learning, where it theoretically justifies some recently proposed constructions.

The result is then extended to Cartesian products, yielding universal approximation theorem for tree-structured domains, which naturally occur in data-exchange

formats like JSON, XML, YAML, AVRO, and ProtoBuffer. This has important practical implications, as it enables to automatically create an architecture of neural networks for processing structured data (AutoML paradigms), as demonstrated by an accompanied library for JSON format.

\*\*\*\*\*

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

Yuping Luo, Huazhe Xu, Yuanzhi Li, Yuandong Tian, Trevor Darrell, Tengyu Ma

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.

\*\*\*\*\*

#### Coupled Recurrent Models for Polyphonic Music Composition

John Thickstun, Zaid Harchaoui, Dean P. Foster, Sham M. Kakade

This work describes a novel recurrent model for music composition, which accounts for the rich statistical structure of polyphonic music. There are many ways to factor the probability distribution over musical scores; we consider the merits of various approaches and propose a new factorization that decomposes a score into a collection of concurrent, coupled time series: "parts." The model we propose borrows ideas from both convolutional neural models and recurrent neural models; we argue that these ideas are natural for capturing music's pitch invariances, temporal structure, and polyphony.

We train generative models for homophonic and polyphonic composition on the Kern Scores dataset (Sapp, 2005), a collection of 2,300 musical scores comprised of a round 2.8 million notes spanning time from the Renaissance to the early 20th century. While evaluation of generative models is known to be hard (Theis et al., 2016), we present careful quantitative results using a unit-adjusted cross entropy metric that is independent of how we factor the distribution over scores. We also present qualitative results using a blind discrimination test.

\*\*\*\*\*

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

Josue Nassar, Scott Linderman, Monica Bugallo, Il Memming Park

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 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 exa

mples, we show how these models outperform existing methods in both interpretability and predictive capability.

\*\*\*\*\*

Reinforced Pipeline Optimization: Behaving Optimally with Non-Differentiables

Aijun Bai, Dongdong Chen, Gang Hua, Lu Yuan

Many machine learning systems are implemented as pipelines. A pipeline is essentially a chain/network of information processing units. As information flows in and out and gradients vice versa, ideally, a pipeline can be trained end-to-end via backpropagation provided with the right supervision and loss function. However, this is usually impossible in practice, because either the loss function itself may be non-differentiable, or there may exist some non-differentiable units. One popular way to superficially resolve this issue is to separate a pipeline into a set of differentiable sub-pipelines and train them with isolated loss functions. Yet, from a decision-theoretical point of view, this is equivalent to making myopic decisions using ad hoc heuristics along the pipeline while ignoring the real utility, which prevents the pipeline from behaving optimally. In this paper, we show that by converting a pipeline into a stochastic counterpart, it can then be trained end-to-end in the presence of non-differentiable parts. Thus, the resulting pipeline is optimal under certain conditions with respect to any criterion attached to it. In experiments, we apply the proposed approach - reinforced pipeline optimization - to Faster R-CNN, a state-of-the-art object detection pipeline, and obtain empirically near-optimal object detectors consistent with its base design in terms of mean average precision.

\*\*\*\*\*

GenEval: A Benchmark Suite for Evaluating Generative Models

Anton Bakhtin, Arthur Szlam, Marc'Aurelio Ranzato

Generative models are important for several practical applications, from low level image processing tasks, to model-based planning in robotics. More generally, the study of generative models is motivated by the long-standing endeavor to model uncertainty and to discover structure by leveraging unlabeled data. Unfortunately, the lack of an ultimate task of interest has hindered progress in the field, as there is no established way to compare models and, often times, evaluation is based on mere visual inspection of samples drawn from such models.

In this work, we aim at addressing this problem by introducing a new benchmark evaluation suite, dubbed \textit{GenEval}.

GenEval hosts a large array of distributions capturing many important properties of real datasets, yet in a controlled setting, such as lower intrinsic dimensionality, multi-modality, compositionality, independence and causal structure. Any model can be easily plugged for evaluation, provided it can generate samples.

Our extensive evaluation suggests that different models have different strengths, and that GenEval is a great tool to gain insights about how models and metrics work.

We offer GenEval to the community~\footnote{Available at: \it{coming soon}.} and believe that this benchmark will facilitate comparison and development of new generative models.

\*\*\*\*\*

Diagnosing and Enhancing VAE Models

Bin Dai, David Wipf

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 hyp

erparameters 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}](https://github.com/daib13/TwoStageVAE).

\*\*\*\*\*

Efficiently testing local optimality and escaping saddles for ReLU networks

Chulhee Yun, Suvrit Sra, Ali Jadbabaie

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

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.

\*\*\*\*\*

Machine Translation With Weakly Paired Bilingual Documents

Lijun Wu, Jinhua Zhu, Di He, Fei Gao, Xu Tan, Tao Qin, Tie-Yan Liu

Neural machine translation, which achieves near human-level performance in some languages, strongly relies on the availability of large amounts of parallel sentences, which hinders its applicability to low-resource language pairs. Recent works explore the possibility of unsupervised machine translation with monolingual data only, leading to much lower accuracy compared with the supervised one. Observing that weakly paired bilingual documents are much easier to collect than bilingual sentences, e.g., from Wikipedia, news websites or books, in this paper, we investigate the training of translation models with weakly paired bilingual documents. Our approach contains two components/steps. First, we provide a simple approach to mine implicitly bilingual sentence pairs from document pairs which can then be used as supervised signals for training. Second, we leverage the topic consistency of two weakly paired documents and learn the sentence-to-sentence translation by constraining the word distribution-level alignments. We evaluate our proposed method on weakly paired documents from Wikipedia on four tasks, the widely used WMT16 German  $\rightarrow$  English and WMT13 Spanish  $\rightarrow$  English tasks, and obtain \$24.1\$/\$30.3\$ and \$28.0\$/\$27.6\$ BLEU points separately, outperforming

state-of-the-art unsupervised results by more than 5 BLEU points and reducing the gap between unsupervised translation and supervised translation up to 50\%.

\*\*\*\*\*

#### An Active Learning Framework for Efficient Robust Policy Search

Sai Kiran Narayanaswami, Nandan Sudarsanam, Balaraman Ravindran

Robust Policy Search is the problem of learning policies that do not degrade in performance when subject to unseen environment model parameters. It is particularly relevant for transferring policies learned in a simulation environment to the real world. Several existing approaches involve sampling large batches of trajectories which reflect the differences in various possible environments, and then selecting some subset of these to learn robust policies, such as the ones that result in the worst performance. We propose an active learning based framework, EffActS, to selectively choose model parameters for this purpose so as to collect only as much data as necessary to select such a subset. We apply this framework to an existing method, namely EPOpt, and experimentally validate the gains in sample efficiency and the performance of our approach on standard continuous control tasks. We also present a Multi-Task Learning perspective to the problem of Robust Policy Search, and draw connections from our proposed framework to existing work on Multi-Task Learning.

\*\*\*\*\*

#### Diverse Machine Translation with a Single Multinomial Latent Variable

Tianxiao Shen, Myle Ott, Michael Auli, Marc'Aurelio Ranzato

There are many ways to translate a sentence into another language. Explicit modeling of such uncertainty may enable better model fitting to the data and it may enable users to express a preference for how to translate a piece of content. Latent variable models are a natural way to represent uncertainty. Prior work investigated the use of multivariate continuous and discrete latent variables, but their interpretation and use for generating a diverse set of hypotheses have been elusive. In this work, we drastically simplify the model, using just a single multinomial latent variable. The resulting mixture of experts model can be trained efficiently via hard-EM and can generate a diverse set of hypothesis by parallel greedy decoding. We perform extensive experiments on three WMT benchmark data sets that have multiple human references, and we show that our model provides a better trade-off between quality and diversity of generations compared to all baseline methods. \footnote{Code to reproduce this work is available at: [anonymized URL](#).}

\*\*\*\*\*

#### Incremental Few-Shot Learning with Attention Attractor Networks

Mengye Ren, Renjie Liao, Ethan Fetaya, Richard S. Zemel

Machine learning classifiers are often trained to recognize a set of pre-defined classes. However, in many real applications, it is often desirable to have the flexibility of learning additional concepts, without re-training on the full training set. This paper addresses this problem, incremental few-shot learning, where a regular classification network has already been trained to recognize a set of base classes; and several extra novel classes are being considered, each with only a few labeled examples. After learning the novel classes, the model is then evaluated on the overall performance of both base and novel classes. To this end, we propose a meta-learning model, the Attention Attractor Network, which regularizes the learning of novel classes. In each episode, we train a set of new weights to recognize novel classes until they converge, and we show that the technique of recurrent back-propagation can back-propagate through the optimization process and facilitate the learning of the attractor network regularizer. We demonstrate that

t the learned

attractor network can recognize novel classes while remembering old classes without the need to

review the original training set, outperforming baselines that do not rely on an iterative

optimization process.

\*\*\*\*\*

Visualizing and Understanding the Semantics of Embedding Spaces via Algebraic Formulae

Piero Molino, Yang Wang, Jiawei Zhang

Embeddings are a fundamental component of many modern machine learning and natural language processing models.

Understanding them and visualizing them is essential for gathering insights about the information they capture and the behavior of the models.

State of the art in analyzing embeddings consists in projecting them in two-dimensional planes without any interpretable semantics associated to the axes of the projection, which makes detailed analyses and comparison among multiple sets of embeddings challenging.

In this work, we propose to use explicit axes defined as algebraic formulae over embeddings to project them into a lower dimensional, but semantically meaningful subspace, as a simple yet effective analysis and visualization methodology.

This methodology assigns an interpretable semantics to the measures of variability and the axes of visualizations, allowing for both comparisons among different sets of embeddings and fine-grained inspection of the embedding spaces.

We demonstrate the power of the proposed methodology through a series of case studies that make use of visualizations constructed around the underlying methodology and through a user study. The results show how the methodology is effective at providing more profound insights than classical projection methods and how it is widely applicable to many other use cases.


\*\*\*\*\*

Pearl: Prototype lEarning via Rule Lists

Tianfan Fu\*, Tian Gao\*, Cao Xiao\*, Tengfei Ma\*, Jimeng Sun

Deep neural networks have demonstrated promising prediction and classification performance on many healthcare applications. However, the interpretability of those models are often lacking. On the other hand, classical interpretable models such as rule lists or decision trees do not lead to the same level of accuracy as

deep neural networks and can often be too complex to interpret (due to the potentially large depth of rule lists). In this work, we present PEARL, Prototype lEarning via Rule Lists, which iteratively uses rule lists to guide a neural network to learn representative data prototypes. The resulting prototype neural network provides accurate prediction, and the prediction can be easily explained by

prototype and its guiding rule lists. Thanks to the prediction power of neural networks, the rule lists from  prototypes are more concise and hence provide better interpretability. On two real-world electronic healthcare records (EHR) datasets, PEARL consistently outperforms all baselines across both datasets, especially achieving performance improvement over conventional rule learning by up to 28% and over prototype learning by up to 3%. Experimental results also show that the resulting interpretation of PEARL is simpler than the standard rule learning.

\*\*\*\*\*

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

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

e 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 failure rates of agents multiple orders of magnitude faster than standard evaluation schemes, in minutes to hours rather than days.

\*\*\*\*\*

TopicGAN: Unsupervised Text Generation from Explainable Latent Topics

Yau-Shian Wang, Yun-Nung Chen, Hung-Yi Lee

Learning discrete representations of data and then generating data from the discovered representations have been increasingly studied because the obtained discrete representations can benefit unsupervised learning. However, the performance of learning discrete representations of textual data with deep generative models has not been widely explored. In addition, although generative adversarial networks (GAN) have shown impressive results in many areas such as image generation, for text generation, it is notorious for extremely difficult to train. In this work, we propose TopicGAN, a two-step text generative model, which is able to solve those two important problems simultaneously. In the first step, it discovers the latent topics and produced bag-of-words according to the latent topics. In the second step, it generates text from the produced bag-of-words. In our experiments, we show our model can discover meaningful discrete latent topics of texts in an unsupervised fashion and generate high quality natural language from the discovered latent topics.

\*\*\*\*\*

Competitive experience replay

Hao Liu, Alexander Trott, Richard Socher, Caiming Xiong

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 convergence and improved task performance.

\*\*\*\*\*

A Max-Affine Spline Perspective of Recurrent Neural Networks

Zichao Wang, Randall Balestriero, Richard Baraniuk

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

t 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.

\*\*\*\*\*

A Differentiable Self-disambiguated Sense Embedding Model via Scaled Gumbel Softmax

Fenfei Guo, Mohit Iyyer, Leah Findlater, Jordan Boyd-Graber

We present a differentiable multi-prototype word representation model that disentangles senses of polysemous words and produces meaningful sense-specific embeddings without external resources. It jointly learns how to disambiguate senses given local context and how to represent senses using hard attention. Unlike previous multi-prototype models, our model approximates discrete sense selection in a differentiable manner via a modified Gumbel softmax. We also propose a novel human evaluation task that quantitatively measures (1) how meaningful the learned sense groups are to humans and (2) how well the model is able to disambiguate senses given a context sentence. Our model outperforms competing approaches on both human evaluations and multiple word similarity tasks.

\*\*\*\*\*

PIE: Pseudo-Invertible Encoder

Jan Jetze Beitler, Ivan Sosnovik, Arnold Smeulders

We consider the problem of information compression from high dimensional data. Where many studies consider the problem of compression by non-invertible transformations, we emphasize the importance of invertible compression. We introduce a new class of likelihood-based auto encoders with pseudo bijective architecture, which we call Pseudo Invertible Encoders. We provide the theoretical explanation of their principles. We evaluate Gaussian Pseudo Invertible Encoder on MNIST, where our model outperforms WAE and VAE in sharpness of the generated images.

\*\*\*\*\*

Probabilistic Knowledge Graph Embeddings

Farnood Salehi, Robert Bamler, Stephan Mandt

We develop a probabilistic extension of state-of-the-art embedding models for link prediction in relational knowledge graphs. Knowledge graphs are collections of relational facts, where each fact states that a certain relation holds between two entities, such as people, places, or objects. We argue that knowledge graphs should be treated within a Bayesian framework because even large knowledge graphs typically contain only few facts per entity, leading effectively to a small data problem where parameter uncertainty matters. We introduce a probabilistic reinterpretation of the DistMult (Yang et al., 2015) and ComplEx (Trouillon et al., 2016) models and employ variational inference to estimate a lower bound on the marginal likelihood of the data. We find that the main benefit of the Bayesian approach is that it allows for efficient, gradient based optimization over hyperparameters, which would lead to divergences in a non-Bayesian treatment. Models with such learned hyperparameters improve over the state-of-the-art by a significant margin, as we demonstrate on several benchmarks.

\*\*\*\*\*

Cross-Task Knowledge Transfer for Visually-Grounded Navigation

Devendra Singh Chaplot, Lisa Lee, Ruslan Salakhutdinov, Devi Parikh, Dhruv Batra

Recent efforts on training visual navigation agents conditioned on language using deep reinforcement learning have been successful in learning policies for two different tasks: learning to follow navigational instructions and embodied question answering. In this paper, we aim to learn a multitask model capable of jointly learning both tasks, and transferring knowledge of words and their grounding



in visual objects across tasks. The proposed model uses a novel Dual-Attention unit to disentangle the knowledge of words in the textual representations and visual objects in the visual representations, and align them with each other. This disentangled task-invariant alignment of representations facilitates grounding and knowledge transfer across both tasks. We show that the proposed model outperforms a range of baselines on both tasks in simulated 3D environments. We also show that this disentanglement of representations makes our model modular, interpretable, and allows for zero-shot transfer to instructions containing new words by leveraging object detectors.

\*\*\*\*\*

Manifold regularization with GANs for semi-supervised learning

Bruno Lecouat, Chuan-Sheng Foo, Houssam Zenati, Vijay Chandrasekhar

Generative Adversarial Networks are powerful generative models that can model the manifold of natural images. We leverage this property to perform manifold regularization by approximating a variant of the Laplacian norm using a Monte Carlo approximation that is easily computed with the GAN. When incorporated into the semi-supervised feature-matching GAN we achieve state-of-the-art results for semi-supervised learning on CIFAR-10 benchmarks when few labels are used, with a method that is significantly easier to implement than competing methods. We find that manifold regularization improves the quality of generated images, and is affected by the quality of the GAN used to approximate the regularizer.

\*\*\*\*\*

Top-Down Neural Model For Formulae

Karel Chvalovský

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.

\*\*\*\*\*

TarMAC: Targeted Multi-Agent Communication

Abhishek Das, Theophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, Joelle Pineau

We explore the collaborative multi-agent setting where a team of deep reinforcement learning agents attempt to solve a shared task in partially observable environments. In this scenario, learning an effective communication protocol is key. We propose a communication protocol that allows for targeted communication, where agents learn *\emph{what}* messages to send and *\emph{who}* to send them to. Additionally, we introduce a multi-stage communication approach where the agents coordinate via several rounds of communication before taking an action in the environment. We evaluate our approach on several cooperative multi-agent tasks, of varying difficulties with varying number of agents, in a variety of environments ranging from 2D grid layouts of shapes and simulated traffic junctions to complex 3D indoor environments. We demonstrate the benefits of targeted as well as multi-stage communication. Moreover, we show that the targeted communication strategies learned by the agents are quite interpretable and intuitive.

\*\*\*\*\*

Feature-Wise Bias Amplification

Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, Anupam Datta

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 while some bias can

not 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.

\*\*\*\*\*

Knows When it Doesn't Know: Deep Abstaining Classifiers

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeffrey Bilmes, Gopinath Chennupati, Jamal Mohd-Yusof

We introduce the deep abstaining classifier -- a deep neural network trained with a novel loss function that provides an abstention option during training. This allows the DNN to abstain on confusing or difficult-to-learn examples while improving performance on the non-abstained samples. We show that such deep abstaining classifiers can: (i) learn representations for structured noise -- where noisy training labels or confusing examples are correlated with underlying features -- and then learn to abstain based on such features; (ii) enable robust learning in the presence of arbitrary or unstructured noise by identifying noisy samples; and (iii) be used as an effective out-of-category detector that learns to reliably abstain when presented with samples from unknown classes. We provide analytical results on loss function behavior that enable automatic tuning of accuracy and coverage, and demonstrate the utility of the deep abstaining classifier using multiple image benchmarks. Results indicate significant improvement in learning in the presence of label noise.

\*\*\*\*\*

Sorting out Lipschitz function approximation

Cem Anil, James Lucas, Roger B. Grosse

Training neural networks subject to a Lipschitz constraint is useful for generalization bounds, provable adversarial robustness, interpretable gradients, and Wasserstein distance estimation. By the composition property of Lipschitz functions, it suffices to ensure that each individual affine transformation or nonlinear activation function is 1-Lipschitz. The challenge is to do this while maintaining the expressive power. We identify a necessary property for such an architecture: each of the layers must preserve the gradient norm during backpropagation. Based on this, we propose to combine a gradient norm preserving activation function, GroupSort, with norm-constrained weight matrices. We show that norm-constrained GroupSort architectures are universal Lipschitz function approximators. Empirically, we show that norm-constrained GroupSort networks achieve tighter estimates of Wasserstein distance than their ReLU counterparts and can achieve provable adversarial robustness guarantees with little cost to accuracy.

\*\*\*\*\*

On Difficulties of Probability Distillation

Chin-Wei Huang, Faruk Ahmed, Kundan Kumar, Alexandre Lacoste, Aaron Courville

Probability distillation has recently been of interest to deep learning practitioners as it presents a practical solution for sampling from autoregressive models for deployment in real-time applications. We identify a pathological optimization issue with the commonly adopted stochastic minimization of the (reverse) KL divergence, owing to sparse gradient signal from the teacher model due to curse of dimensionality. We also explore alternative principles for distillation, and show that one can achieve qualitatively better results than with KL minimization.

.

\*\*\*\*\*

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

Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, Yizhou Wang

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 me

thod 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/zfwl226/active\\_tracking\\_rl](https://github.com/zfwl226/active_tracking_rl)

\*\*\*\*\*

#### Theoretical and Empirical Study of Adversarial Examples

Fuchen Liu, Hongwei Shang, Hong Zhang

Many techniques are developed to defend against adversarial examples at scale. So far, the most successful defenses generate adversarial examples during each training step and add them to the training data. Yet, this brings significant computational overhead. In this paper, we investigate defenses against adversarial attacks. First, we propose feature smoothing, a simple data augmentation method with little computational overhead. Essentially, feature smoothing trains a neural network on virtual training data as an interpolation of features from a pair of samples, with the new label remaining the same as the dominant data point. The intuition behind feature smoothing is to generate virtual data points as close as adversarial examples, and to avoid the computational burden of generating data during training. Our experiments on MNIST and CIFAR10 datasets explore different combinations of known regularization and data augmentation methods and show that feature smoothing with logit squeezing performs best for both adversarial and clean accuracy. Second, we propose an unified framework to understand the connections and differences among different efficient methods by analyzing the biases and variances of decision boundary. We show that under some symmetrical assumptions, label smoothing, logit squeezing, weight decay, mix up and feature smoothing all produce an unbiased estimation of the decision boundary with smaller estimated variance. All of those methods except weight decay are also stable when the assumptions no longer hold.

\*\*\*\*\*

#### Causal Reasoning from Meta-reinforcement learning

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, Zeb Kurth-Nelson

Discovering and exploiting the causal structure in the environment is a crucial challenge for intelligent agents. Here we explore whether modern deep reinforcement learning can be used to train agents to perform causal reasoning. We adopt a meta-learning approach, where the agent learns a policy for conducting experiments via causal interventions, in order to support a subsequent task which rewards making accurate causal inferences. We also found the agent could make sophisticated counterfactual predictions, as well as learn to draw causal inferences from purely observational data. Though powerful formalisms for causal reasoning have been developed, applying them in real-world domains can be difficult because fitting to large amounts of high dimensional data often requires making idealized assumptions. Our results suggest that causal reasoning in complex settings may benefit from powerful learning-based approaches. More generally, this work may offer new strategies for structured exploration in reinforcement learning, by providing agents with the ability to perform—and interpret—experiments.

\*\*\*\*\*

## The Conditional Entropy Bottleneck

Ian Fischer

We present a new family of objective functions, which we term the Conditional Entropy Bottleneck (CEB). These objectives are motivated by the Minimum Necessary Information (MNI) criterion. We demonstrate the application of CEB to classification tasks. We show that CEB gives: well-calibrated predictions; strong detection of challenging out-of-distribution examples and powerful whitebox adversarial examples; and substantial robustness to those adversaries. Finally, we report that CEB fails to learn from information-free datasets, providing a possible resolution to the problem of generalization observed in Zhang et al. (2016).

\*\*\*\*\*

## Learning to Learn with Conditional Class Dependencies

Xiang Jiang, Mohammad Havaei, Farshid Varno, Gabriel Chartrand, Nicolas Chapados, Stan Matwin

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.

\*\*\*\*\*

## Open Loop Hyperparameter Optimization and Determinantal Point Processes

Jesse Dodge, Kevin Jamieson, Noah Smith

Driven by the need for parallelizable hyperparameter optimization methods, this paper studies open loop search methods: sequences that are predetermined and can be generated before a single configuration is evaluated. Examples include grid search, uniform random search, low discrepancy sequences, and other sampling distributions.

In particular, we propose the use of k-determinantal point processes in hyperparameter optimization via random search. Compared to conventional uniform random search where hyperparameter settings are sampled independently, a k-DPP promotes diversity. We describe an approach that transforms hyperparameter search spaces for efficient use with a k-DPP. In addition, we introduce a novel Metropolis-Hastings algorithm which can sample from k-DPPs defined over any space from which uniform samples can be drawn, including spaces with a mixture of discrete and continuous dimensions or tree structure. Our experiments show significant benefits in realistic scenarios with a limited budget for training supervised learners, whether in serial or parallel.

\*\*\*\*\*

## Self-Supervised Generalisation with Meta Auxiliary Learning

Shikun Liu, Edward Johns, Andrew Davison

Auxiliary learning has been shown to improve the generalisation performance of a principal task. But typically, this requires manually-defined auxiliary tasks based on domain knowledge. In this paper, we consider that it may be possible to automatically learn these auxiliary tasks to best suit the principal task, towards optimum auxiliary tasks without any human knowledge. We propose a novel method, Meta Auxiliary Learning (MAXL), which we design for the task of image classification, where the auxiliary task is hierarchical sub-class image classification. The role of the meta learner is to determine sub-class target labels to train a multi-task evaluator, such that these labels improve the generalisation performance on the principal task. Experiments on three different CIFAR datasets show that MAXL outperforms baseline auxiliary learning methods, and is competitive ev

en with a method which uses human-defined sub-class hierarchies. MAXL is self-supervised and general, and therefore offers a promising new direction towards automated generalisation.

\*\*\*\*\*

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

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

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.

\*\*\*\*\*

Can I trust you more? Model-Agnostic Hierarchical Explanations

Michael Tsang, Youbang Sun, Dongxu Ren, Beibei Xin, Yan Liu

Interactions such as double negation in sentences and scene interactions in images are common forms of complex dependencies captured by state-of-the-art machine learning models. We propose Mahé, a novel approach to provide Model-Agnostic Hierarchical Explanations of how powerful machine learning models, such as deep neural networks, capture these interactions as either dependent on or free of the context of data instances. Specifically, Mahé provides context-dependent explanations by a novel local interpretation algorithm that effectively captures any-order interactions, and obtains context-free explanations through generalizing context-dependent interactions to explain global behaviors. Experimental results show that Mahé obtains improved local interaction interpretations over state-of-the-art methods and successfully provides explanations of interactions that are co

ntext-free.

\*\*\*\*\*

#### A Mean Field Theory of Batch Normalization

Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, Samuel S. Schoenholz

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.

\*\*\*\*\*

#### Analyzing Federated Learning through an Adversarial Lens

Arjun Nitin Bhagoji, Supriyo Chakraborty, Seraphin Calo, Prateek Mittal

Federated learning distributes model training among a multitude of agents, who, guided by privacy concerns, perform training using their local data but share only model parameter updates, for iterative aggregation at the server. In this work, we explore the threat of model poisoning attacks on federated learning initiated by a single, non-colluding malicious agent where the adversarial objective is to cause the model to misclassify a set of chosen inputs with high confidence. We explore a number of strategies to carry out this attack, starting with simple boosting of the malicious agent's update to overcome the effects of other agents' updates. To increase attack stealth, we propose an alternating minimization strategy, which alternately optimizes for the training loss and the adversarial objective. We follow up by using parameter estimation for the benign agents' updates to improve on attack success. Finally, we use a suite of interpretability techniques to generate visual explanations of model decisions for both benign and malicious models and show that the explanations are nearly visually indistinguishable. Our results indicate that even a highly constrained adversary can carry out model poisoning attacks while simultaneously maintaining stealth, thus highlighting the vulnerability of the federated learning setting and the need to develop effective defense strategies.

\*\*\*\*\*

#### COMPOSITION AND DECOMPOSITION OF GANS

Yeu-Chern Harn, Zhenghao Chen, Vladimir Jojic

In this work, we propose a composition/decomposition framework for adversarially training generative models on composed data - data where each sample can be thought of as being constructed from a fixed number of components. In our framework, samples are generated by sampling components from component generators and feeding these components to a composition function which combines them into a "composed sample". This compositional training approach improves the modularity, extensibility and interpretability of Generative Adversarial Networks (GANs) - providing a principled way to incrementally construct complex models out of simpler component models, and allowing for explicit "division of responsibility" between these components. Using this framework, we define a family of learning tasks and evaluate their feasibility on two datasets in two different data modalities (image and text). Lastly, we derive sufficient conditions such that these compositional generative models are identifiable. Our work provides a principled approach to building on pretrained generative models or for exploiting the compositional

nature of data distributions to train extensible and interpretable models.

\*\*\*\*\*

#### Learning Backpropagation-Free Deep Architectures with Kernels

Shiyu Duan, Shujian Yu, Yunmei Chen, Jose Principe

One can substitute each neuron in any neural network with a kernel machine and obtain a counterpart powered by kernel machines. The new network inherits the expressive power and architecture of the original but works in a more intuitive way since each node enjoys the simple interpretation as a hyperplane (in a reproducing kernel Hilbert space). Further, using the kernel multilayer perceptron as an example, we prove that in classification, an optimal representation that minimizes the risk of the network can be characterized for each hidden layer. This result removes the need of backpropagation in learning the model and can be generalized to any feedforward kernel network. Moreover, unlike backpropagation, which turns models into black boxes, the optimal hidden representation enjoys an intuitive geometric interpretation, making the dynamics of learning in a deep kernel network simple to understand. Empirical results are provided to validate our theory.

\*\*\*\*\*

#### Unsupervised one-to-many image translation

Samuel Lavoie-Marchildon, Sebastien Lachapelle, Mikołaj Bińkowski, Aaron Courville, Yoshua Bengio, R Devon Hjelm

We perform completely unsupervised one-sided image to image translation between a source domain  $X$  and a target domain  $Y$  such that we preserve relevant underlying shared semantics (e.g., class, size, shape, etc).

In particular, we are interested in a more difficult case than those typically addressed in the literature, where the source and target are "far" enough that reconstruction-style or pixel-wise approaches fail.

We argue that transferring (i.e., *translating*) said relevant information should involve both discarding source domain-specific information while incorporate target domain-specific information, the latter of which we model with a noisy prior distribution.

In order to avoid the degenerate case where the generated samples are only explained by the prior distribution, we propose to minimize an estimate of the mutual information between the generated sample and the sample from the prior distribution. We discover that the architectural choices are an important factor to consider in order to preserve the shared semantic between  $X$  and  $Y$ .

We show state of the art results on the MNIST to SVHN task for unsupervised image to image translation.

\*\*\*\*\*

#### A Closer Look at Few-shot Classification

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, Jia-Bin Huang

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.

\*\*\*\*\*

#### Looking for ELMo's friends: Sentence-Level Pretraining Beyond Language Modeling

Samuel R. Bowman, Ellie Pavlick, Edouard Grave, Benjamin Van Durme, Alex Wang, Jan Hui, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen

Work on the problem of contextualized word representation—the development of reusable neural network components for sentence understanding—has recently seen a surge of progress centered on the unsupervised pretraining task of language modeling with methods like ELMo (Peters et al., 2018). This paper contributes the first large-scale systematic study comparing different pretraining tasks in this context, both as complements to language modeling and as potential alternatives. The primary results of the study support the use of language modeling as a pretraining task and set a new state of the art among comparable models using multitask learning with language models. However, a closer look at these results reveals worryingly strong baselines and strikingly varied results across target tasks, suggesting that the widely-used paradigm of pretraining and freezing sentence encoders may not be an ideal platform for further work.

\*\*\*\*\*

#### Variational Domain Adaptation

Hirono Okamoto, Shohei Ohsawa, Itto Higuchi, Haruka Murakami, Mizuki Sango, Zhenghang Cui, Masahiro Suzuki, Hiroshi Kajino, Yutaka Matsuo

This paper proposes variational domain adaptation, a unified, scalable, simple framework for learning multiple distributions through variational inference. Unlike the existing methods on domain transfer through deep generative models, such as StarGAN (Choi et al., 2017) and UFDN (Liu et al., 2018), the variational domain adaptation has three advantages. Firstly, the samples from the target are not required. Instead, the framework requires one known source as a prior  $p(x)$  and binary discriminators,  $p(\mathcal{D}_i|x)$ , discriminating the target domain  $\mathcal{D}_i$  from others. Consequently, the framework regards a target as a posterior that can be explicitly formulated through the Bayesian inference,  $p(x|\mathcal{D}_i) \propto p(\mathcal{D}_i|x)p(x)$ , as exhibited by a further proposed model of dual variational autoencoder (DualVAE). Secondly, the framework is scalable to large-scale domains. As well as VAE encodes a sample  $x$  as a mode on a latent space:  $\mu(x) \in \mathcal{Z}$ , DualVAE encodes a domain  $\mathcal{D}_i$  as a mode on the dual latent space  $\mu^*(\mathcal{D}_i) \in \mathcal{Z}^*$ , named domain embedding. It reformulates the posterior with a natural pairing  $\langle \cdot, \cdot \rangle: \mathcal{Z} \times \mathcal{Z}^* \rightarrow \mathbb{R}$ , which can be expanded to uncountable infinite domains such as continuous domains as well as interpolation. Thirdly, DualVAE fastly converges without sophisticated automatic/manual hyperparameter search in comparison to GANs as it requires only one additional parameter to VAE. Through the numerical experiment, we demonstrate the three benefits with multi-domain image generation task on CelebA with up to 60 domains, and exhibits that DualVAE records the state-of-the-art performance outperforming StarGAN and UFDN.

\*\*\*\*\*

#### Fast Exploration with Simplified Models and Approximately Optimistic Planning in Model Based Reinforcement Learning

Ramtin Keramati, Jay Whang, Patrick Cho, Emma Brunskill

Humans learn to play video games significantly faster than the state-of-the-art reinforcement learning (RL) algorithms. People seem to build simple models that are easy to learn to support planning and strategic exploration. Inspired by this, we investigate two issues in leveraging model-based RL for sample efficiency. First we investigate how to perform strategic exploration when exact planning is not feasible and empirically show that optimistic Monte Carlo Tree Search outperforms posterior sampling methods. Second we show how to learn simple deterministic models to support fast learning using object representation. We illustrate the benefit of these ideas by introducing a novel algorithm, Strategic Object Oriented Reinforcement Learning (SOORL), that outperforms state-of-the-art algorithms in the game of Pitfall! in less than 50 episodes.

\*\*\*\*\*

#### STCN: Stochastic Temporal Convolutional Networks



Emre Aksan,Otmar Hilliges

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.

\*\*\*\*\*

Select Via Proxy: Efficient Data Selection For Training Deep Networks

Cody Coleman,Stephen Mussmann,Baharan Mirzasoleiman,Peter Bailis,Percy Liang,Jurgen Leskovec,Matei Zaharia

At internet scale, applications collect a tremendous amount of data by logging user events, analyzing text, and collecting images. This data powers a variety of machine learning models for tasks such as image classification, language modeling, content recommendation, and advertising. However, training large models over all available data can be computationally expensive, creating a bottleneck in the development of new machine learning models. In this work, we develop a novel approach to efficiently select a subset of training data to achieve faster training with no loss in model predictive performance. In our approach, we first train a small proxy model quickly, which we then use to estimate the utility of individual training data points, and then select the most informative ones for training the large target model. Extensive experiments show that our approach leads to a 1.6x and 1.8x speed-up on CIFAR10 and SVHN by selecting 60% and 50% subsets of the data, while maintaining the predictive performance of the model trained on the entire dataset.

\*\*\*\*\*

Successor Options : An Option Discovery Algorithm for Reinforcement Learning

Manan Tomar\*,Rahul Ramesh\*,Balaraman Ravindran

Hierarchical Reinforcement Learning is a popular method to exploit temporal abstractions in order to tackle the curse of dimensionality. The options framework is one such hierarchical framework that models the notion of skills or options. However, learning a collection of task-agnostic transferable skills is a challenging task. Option discovery typically entails using heuristics, the majority of which revolve around discovering bottleneck states. In this work, we adopt a method complementary to the idea of discovering bottlenecks. Instead, we attempt to discover "landmark" sub-goals which are prototypical states of well connected regions. These sub-goals are points from which densely connected set of states are easily accessible. We propose a new model called Successor options that leverages Successor Representations to achieve the same. We also design a novel pseudo-reward for learning the intra-option policies. Additionally, we describe an Incremental Successor options model that iteratively builds options and explores in environments where exploration through primitive actions is inadequate to form the Successor Representations. Finally, we demonstrate the efficacy of our approach on a collection of grid worlds and on complex high dimensional environments like Deepmind-Lab.

\*\*\*\*\*

Multi-Objective Value Iteration with Parameterized Threshold-Based Safety Constraints

Hussein Sibai,Sayan Mitra

We consider an environment with multiple reward functions. One of them represent

s goal achievement and the others represent instantaneous safety conditions. We consider a scenario where the safety rewards should always be above some thresholds. The thresholds are parameters with values that differ between users.

%The thresholds are not known at the time the policy is being designed.

We efficiently compute a family of policies that cover all threshold-based constraints and maximize the goal achievement reward. We introduce a new parameterized threshold-based scalarization method of the reward vector that encodes our objective. We present novel data structures to store the value functions of the Bellman equation that allow their efficient computation using the value iteration algorithm. We present results for both discrete and continuous state spaces.

\*\*\*\*\*

Augment your batch: better training with larger batches

Elad Hoffer, Itay Hubara, Niv Giladi, Daniel Soudry

Recently, there is regained interest in large batch training of neural networks, both of theory and practice. New insights and methods allowed certain models to be trained using large batches with no adverse impact on performance. Most works focused on accelerating wall clock training time by modifying the learning rate schedule, without introducing accuracy degradation.

We propose to use large batch training to boost accuracy and accelerate convergence by combining it with data augmentation. Our method, "batch augmentation", suggests using multiple instances of each sample at the same large batch. We show empirically that this simple yet effective method improves convergence and final generalization accuracy. We further suggest possible reasons for its success.

\*\*\*\*\*

Adaptive Sample-space & Adaptive Probability coding: a neural-network based approach for compression

Ken Nakanishi, Shin-ichi Maeda, Takeru Miyato, Masanori Koyama

We propose Adaptive Sample-space & Adaptive Probability (ASAP) coding, an efficient neural-network based method for lossy data compression.

Our ASAP coding distinguishes itself from the conventional method based on adaptive arithmetic coding in that it models the probability distribution for the quantization process in such a way that one can conduct back-propagation for the quantization width that determines the support of the distribution.

Our ASAP also trains the model with a novel, hyper-parameter free multiplicative loss for the rate-distortion tradeoff.

With our ASAP encoder, we are able to compress the image files in the Kodak data set to as low as one fifth the size of the JPEG-compressed image without compromising their visual quality, and achieved the state-of-the-art result in terms of MS-SSIM based rate-distortion tradeoff.

\*\*\*\*\*

Mixed Precision Quantization of ConvNets via Differentiable Neural Architecture Search

Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, Kurt Keutzer

Recent work in network quantization has substantially reduced the time and space complexity of neural network inference, enabling their deployment on embedded and mobile devices with limited computational and memory resources. However, existing quantization methods often represent all weights and activations with the same precision (bit-width). In this paper, we explore a new dimension of the design space: quantizing different layers with different bit-widths. We formulate this problem as a neural architecture search problem and propose a novel differentiable neural architecture search (DNAS) framework to efficiently explore its exponential search space with gradient-based optimization. Experiments show we surpass the state-of-the-art compression of ResNet on CIFAR-10 and ImageNet. Our quantized models with 21.1x smaller model size or 103.9x lower computational cost can still outperform baseline quantized or even full precision models.

\*\*\*\*\*

Uncertainty in Multitask Transfer Learning

Alexandre Lacoste, Boris Oreshkin, Wonchang Chung, Thomas Boquet, Negar Rostamzadeh, David Krueger

Using variational Bayes neural networks, we develop an algorithm capable of accu

mutating knowledge into a prior from multiple different tasks. This results in a rich prior capable of few-shot learning on new tasks. The posterior can go beyond the mean field approximation and yields good uncertainty on the performed experiments. Analysis on toy tasks show that it can learn from significantly different tasks while finding similarities among them. Experiments on Mini-Imagenet reach state of the art with 74.5% accuracy on 5 shot learning. Finally, we provide two new benchmarks, each showing a failure mode of existing meta learning algorithms such as MAML and prototypical Networks.

\*\*\*\*\*

RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space

Zhiqing Sun,Zhi-Hong Deng,Jian-Yun Nie,Jian Tang

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.

\*\*\*\*\*

Q-map: a Convolutional Approach for Goal-Oriented Reinforcement Learning

Fabio Pardo,Vitaly Levдик,Petar Kormushev

Goal-oriented learning has become a core concept in reinforcement learning (RL), extending the reward signal as a sole way to define tasks. However, as parameterizing value functions with goals increases the learning complexity, efficiently reusing past experience to update estimates towards several goals at once becomes desirable but usually requires independent updates per goal.

Considering that a significant number of RL environments can support spatial coordinates as goals, such as on-screen location of the character in ATARI or SNES games, we propose a novel goal-oriented agent called Q-map that utilizes an auto-encoder-like neural network to predict the minimum number of steps towards each coordinate in a single forward pass. This architecture is similar to Horde with parameter sharing and allows the agent to discover correlations between visual patterns and navigation. For example learning how to use a ladder in a game could be transferred to other ladders later.

We show how this network can be efficiently trained with a 3D variant of Q-learning to update the estimates towards all goals at once. While the Q-map agent could be used for a wide range of applications, we propose a novel exploration mechanism in place of epsilon-greedy that relies on goal selection at a desired distance followed by several steps taken towards it, allowing long and coherent exploratory steps in the environment.

We demonstrate the accuracy and generalization qualities of the Q-map agent on a grid-world environment and then demonstrate the efficiency of the proposed exploration mechanism on the notoriously difficult Montezuma's Revenge and Super Mario All-Stars games.

\*\*\*\*\*

A Variational Dirichlet Framework for Out-of-Distribution Detection

Wenhu Chen,Yilin Shen,William Wang,Hongxia Jin

With the recently rapid development in deep learning, deep neural networks have been widely adopted in many real-life applications. However, deep neural networks are also known to have very little control over its uncertainty for test examples, which potentially causes very harmful and annoying consequences in practical scenarios. In this paper, we are particularly interested in designing a higher-order uncertainty metric for deep neural networks and investigate its performance on the out-of-distribution detection task proposed by~\cite{hendrycks2016base

line}. Our method first assumes there exists a underlying higher-order distribution  $\mathcal{P}(z)$ , which generated label-wise distribution  $\mathcal{P}(y)$  over classes on the K-dimension simplex, and then approximate such higher-order distribution via parameterized posterior function  $p_{\theta}(z|x)$  under variational inference framework, finally we use the entropy of learned posterior distribution  $p_{\theta}(z|x)$  as uncertainty measure to detect out-of-distribution examples. However, we identify the overwhelming over-concentration issue in such a framework, which greatly hinders the detection performance. Therefore, we further design a log-smoothing function to alleviate such issue to greatly increase the robustness of the proposed entropy-based uncertainty measure. Through comprehensive experiments on various datasets and architectures, our proposed variational Dirichlet framework with entropy-based uncertainty measure is consistently observed to yield significant improvements over many baseline systems.

\*\*\*\*\*

#### TENSOR RING NETS ADAPTED DEEP MULTI-TASK LEARNING

Xinqi Chen, Ming Hou, Guoxu Zhou, Qibin Zhao

Recent deep multi-task learning (MTL) has been witnessed its success in alleviating data scarcity of some task by utilizing domain-specific knowledge from related tasks. Nonetheless, several major issues of deep MTL, including the effectiveness of sharing mechanisms, the efficiency of model complexity and the flexibility of network architectures, still remain largely unaddressed. To this end, we propose a novel generalized latent-subspace based knowledge sharing mechanism for linking task-specific models, namely tensor ring multi-task learning (TRMTL). TRMTL has a highly compact representation, and it is very effective in transferring task-invariant knowledge while being super flexible in learning task-specific features, successfully mitigating the dilemma of both negative-transfer in lower layers and under-transfer in higher layers. Under our TRMTL, it is feasible for each task to have heterogeneous input data dimensionality or distinct feature sizes at different hidden layers. Experiments on a variety of datasets demonstrate our model is capable of significantly improving each single task's performance, particularly favourable in scenarios where some of the tasks have insufficient data.

\*\*\*\*\*

#### Context Mover's Distance & Barycenters: Optimal transport of contexts for building representations

Sidak Pal Singh, Andreas Hug, Aymeric Dieuleveut, Martin Jaggi

We propose a unified framework for building unsupervised representations of entities and their compositions, by viewing each entity as a histogram (or distribution) over its contexts. This enables us to take advantage of optimal transport and construct representations that effectively harness the geometry of the underlying space containing the contexts. Our method captures uncertainty via modelling the entities as distributions and simultaneously provides interpretability with the optimal transport map, hence giving a novel perspective for building rich and powerful feature representations. As a guiding example, we formulate unsupervised representations for text, and demonstrate it on tasks such as sentence similarity and word entailment detection. Empirical results show strong advantages gained through the proposed framework. This approach can potentially be used for any unsupervised or supervised problem (on text or other modalities) with a co-occurrence structure, such as any sequence data. The key tools at the core of this framework are Wasserstein distances and Wasserstein barycenters.

\*\*\*\*\*

#### Learning to Navigate the Web

Izzeddin Gur, Ulrich Rueckert, Aleksandra Faust, Dilek Hakkani-Tur

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

re 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 on World 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

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

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.

\*\*\*\*\*

#### Neural Predictive Belief Representations

Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo Avila Pires, Rémi Munos

Unsupervised representation learning has succeeded with excellent results in man

y applications. It is an especially powerful tool to learn a good representation of environments with partial or noisy observations. In partially observable domains it is important for the representation to encode a belief state--a sufficient statistic of the observations seen so far. In this paper, we investigate whether it is possible to learn such a belief representation using modern neural architectures. Specifically, we focus on one-step frame prediction and two variants of contrastive predictive coding (CPC) as the objective functions to learn the representations. To evaluate these learned representations, we test how well they can predict various pieces of information about the underlying state of the environment, e.g., position of the agent in a 3D maze. We show that all three methods are able to learn belief representations of the environment--they encode not only the state information, but also its uncertainty, a crucial aspect of belief states. We also find that for CPC multi-step predictions and action-conditioning are critical for accurate belief representations in visually complex environments. The ability of neural representations to capture the belief information has the potential to spur new advances for learning and planning in partially observable domains, where leveraging uncertainty is essential for optimal decision making.

\*\*\*\*\*

#### Understanding Opportunities for Efficiency in Single-image Super Resolution Networks

Royson Lee, Nic Lane, Marko Stankovic, Sourav Bhattacharya

A successful application of convolutional architectures is to increase the resolution of single low-resolution images -- a image restoration task called super-resolution (SR). Naturally, SR is of value to resource constrained devices like mobile phones, electronic photograph frames and televisions to enhance image quality. However, SR demands perhaps the most extreme amounts of memory and compute operations of any mainstream vision task known today, preventing SR from being deployed to devices that require them. In this paper, we perform a early systematic study of system resource efficiency for SR, within the context of a variety of architectural and low-precision approaches originally developed for discriminative neural networks. We present a rich set of insights, representative SR architectures, and efficiency trade-offs; for example, the prioritization of ways to compress models to reach a specific memory and computation target and techniques to compact SR models so that they are suitable for DSPs and FPGAs. As a result of doing so, we manage to achieve better and comparable performance with previous models in the existing literature, highlighting the practicality of using existing efficiency techniques in SR tasks. Collectively, we believe these results provides the foundation for further research into the little explored area of resource efficiency for SR.

\*\*\*\*\*

#### CGNF: Conditional Graph Neural Fields

Tengfei Ma, Cao Xiao, Junyuan Shang, Jimeng Sun

Graph convolutional networks have achieved tremendous success in the tasks of graph node classification. These models could learn a better node representation through encoding the graph structure and node features. However, the correlation between the node labels are not considered. In this paper, we propose a novel architecture for graph node classification, named conditional graph neural fields (CGNF). By integrating the conditional random fields (CRF) in the graph convolutional networks, we explicitly model a joint probability of the entire set of node labels, thus taking advantage of neighborhood label information in the node label prediction task.

Our model could have both the representation capacity of graph neural networks and the prediction power of CRFs. Experiments on several graph datasets demonstrate effectiveness of CGNF.

\*\*\*\*\*

#### Discovering General-Purpose Active Learning Strategies

Ksenia Konyushkova, Raphael Sznitman, Pascal Fua

We propose a general-purpose approach to discovering active learning (AL) strategies from data. These strategies are transferable from one domain to another and

can be used in conjunction with many machine learning models. To this end, we formalize the annotation process as a Markov decision process, design universal state and action spaces and introduce a new reward function that precisely reflects the AL objective of minimizing the annotation cost. We seek to find an optimal (non-myopic) AL strategy using reinforcement learning. We evaluate the learned strategies on multiple unrelated domains and show that they consistently outperform state-of-the-art baselines.

\*\*\*\*\*

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

Dzmitry Bahdanau\*, Shikhar Murty\*, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, Aaron Courville

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

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.

\*\*\*\*\*

Composing Entropic Policies using Divergence Correction

Jonathan J Hunt, Andre Barreto, Timothy P Lillicrap, Nicolas Heess

Deep reinforcement learning (RL) algorithms have made great strides in recent years. An important remaining challenge is the ability to quickly transfer existing skills to novel tasks, and to combine existing skills with newly acquired ones. In domains where tasks are solved by composing skills this capacity holds the promise of dramatically reducing the data requirements of deep RL algorithms, and hence increasing their applicability. Recent work has studied ways of composing behaviors represented in the form of action-value functions. We analyze these methods to highlight their strengths and weaknesses, and point out situations where each of them is susceptible to poor performance. To perform this analysis we extend generalized policy improvement to the max-entropy framework and introduce a method for the practical implementation of successor features in continuous action spaces. Then we propose a novel approach which, in principle, recovers the optimal policy during transfer. This method works by explicitly learning the (discounted, future) divergence between policies. We study this approach in the tabular case and propose a scalable variant that is applicable in multi-dimensional continuous action spaces.

We compare our approach with existing ones on a range of non-trivial continuous

control problems with compositional structure, and demonstrate qualitatively better performance despite not requiring simultaneous observation of all task rewards.

\*\*\*\*\*

#### Optimal Control Via Neural Networks: A Convex Approach

Yize Chen, Yuan Yuan Shi, Baosen Zhang

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.

\*\*\*\*\*

#### Learning to Search Efficient DenseNet with Layer-wise Pruning

Xuanyang Zhang, Hao Liu, Zhanxing Zhu, Zenglin Xu

Deep neural networks have achieved outstanding performance in many real-world applications with the expense of huge computational resources. The DenseNet, one of the recently proposed neural network architecture, has achieved the state-of-the-art performance in many visual tasks. However, it has great redundancy due to the dense connections of the internal structure, which leads to high computational costs in training such dense networks. To address this issue, we design a reinforcement learning framework to search for efficient DenseNet architectures with layer-wise pruning (LWP) for different tasks, while retaining the original advantages of DenseNet, such as feature reuse, short paths, etc. In this framework, an agent evaluates the importance of each connection between any two block layers, and prunes the redundant connections. In addition, a novel reward-shaping trick is introduced to make DenseNet reach a better trade-off between accuracy and float point operations (FLOPs). Our experiments show that DenseNet with LWP is more compact and efficient than existing alternatives.

\*\*\*\*\*

#### Learning to Control Visual Abstractions for Structured Exploration in Deep Reinforcement Learning

Catalin Ionescu, Tejas Kulkarni, Aaron van de Oord, Andriy Mnih, Vlad Mnih

Exploration in environments with sparse rewards is a key challenge for reinforcement learning. How do we design agents with generic inductive biases so that they can explore in a consistent manner instead of just using local exploration schemes like epsilon-greedy? We propose an unsupervised reinforcement learning agent which learns a discrete pixel grouping model that preserves spatial geometry of the sensors and implicitly of the environment as well. We use this representation to derive geometric intrinsic reward functions, like centroid coordinates and area, and learn policies to control each one of them with off-policy learning.

These policies form a basis set of behaviors (options) which allows us explore in a consistent way and use them in a hierarchical reinforcement learning setup to solve for extrinsically defined rewards. We show that our approach can scale to a variety of domains with competitive performance, including navigation in 3D



environments and Atari games with sparse rewards.

\*\*\*\*\*

#### A Proposed Hierarchy of Deep Learning Tasks

Joel Hestness, Sharan Narang, Newsha Ardalani, Heewoo Jun, Hassan Kianinejad, Md. Mostafa Ali Patwary, Yang Yang, Yanqi Zhou, Gregory Diamos, Kenneth Church

As the pace of deep learning innovation accelerates, it becomes increasingly important to organize the space of problems by relative difficulty. Looking to other fields for inspiration, we see analogies to the Chomsky Hierarchy in computational linguistics and time and space complexity in theoretical computer science.

As a complement to prior theoretical work on the data and computational requirements of learning, this paper presents an empirical approach. We introduce a methodology for measuring validation error scaling with data and model size and test tasks in natural language, vision, and speech domains. We find that power-law validation error scaling exists across a breadth of factors and that model size scales sublinearly with data size, suggesting that simple learning theoretic models offer insights into the scaling behavior of realistic deep learning settings, and providing a new perspective on how to organize the space of problems.

We measure the power-law exponent---the "steepness" of the learning curve---and propose using this metric to sort problems by degree of difficulty. There is no data like more data, but some tasks are more effective at taking advantage of more data. Those that are more effective are easier on the proposed scale.

Using this approach, we can observe that studied tasks in speech and vision domains scale faster than those in the natural language domain, offering insight into the observation that progress in these areas has proceeded more rapidly than in natural language.

\*\*\*\*\*

#### Where Off-Policy Deep Reinforcement Learning Fails

Scott Fujimoto, David Meger, Doina Precup

This work examines batch reinforcement learning--the task of maximally exploiting a given batch of off-policy data, without further data collection. We demonstrate that due to errors introduced by extrapolation, standard off-policy deep reinforcement learning algorithms, such as DQN and DDPG, are only capable of learning with data correlated to their current policy, making them ineffective for most off-policy applications. We introduce a novel class of off-policy algorithms, batch-constrained reinforcement learning, which restricts the action space to force the agent towards behaving on-policy with respect to a subset of the given data. We extend this notion to deep reinforcement learning, and to the best of our knowledge, present the first continuous control deep reinforcement learning algorithm which can learn effectively from uncorrelated off-policy data.

\*\*\*\*\*

#### Pix2Scene: Learning Implicit 3D Representations from Images

Sai Rajeswar, Fahim Mannan, Florian Golemo, David Vazquez, Derek Nowrouzezahrai, Aaron Courville

Modelling 3D scenes from 2D images is a long-standing problem in computer vision with implications in, e.g., simulation and robotics. We propose pix2scene, a deep generative-based approach that implicitly models the geometric properties of a scene from images. Our method learns the depth and orientation of scene points visible in images. Our model can then predict the structure of a scene from various, previously unseen view points. It relies on a bi-directional adversarial learning mechanism to generate scene representations from a latent code, inferring the 3D representation of the underlying scene geometry. We showcase a novel differentiable renderer to train the 3D model in an end-to-end fashion, using only images. We demonstrate the generative ability of our model qualitatively on both a custom dataset and on ShapeNet. Finally, we evaluate the effectiveness of the learned 3D scene representation in supporting a 3D spatial reasoning.

\*\*\*\*\*

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

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.

\*\*\*\*\*

Universal discriminative quantum neural networks

Hongxiang Chen,Leonard Wossnig,Hartmut Neven,Simone Severini,Masoud Mohseni  
Quantum mechanics fundamentally forbids deterministic discrimination of quantum states and processes. However, the ability to optimally distinguish various classes of quantum data is an important primitive in quantum information science. In this work, we trained near-term quantum circuits to classify data represented by quantum states using the Adam stochastic optimization algorithm. This is achieved by iterative interactions of a classical device with a quantum processor to discover the parameters of an unknown non-unitary quantum circuit. This circuit learns to simulate the unknown structure of a generalized quantum measurement, or positive-operator valued measure (POVM), that is required to optimally distinguish possible distributions of quantum inputs. Notably we used universal circuit topologies, with a theoretically motivated circuit design which guaranteed that our circuits can perform arbitrary input-output mappings. Our numerical simulations showed that quantum circuits could be trained to discriminate among various pure and mixed quantum states, exhibiting a trade-off between minimizing erroneous and inconclusive outcomes with comparable performance to theoretically optimal POVMs. We trained the circuit on different classes of quantum data and evaluated the generalization error on unseen quantum data. This generalization power hence distinguishes our work from standard circuit optimization and provides an example of quantum machine learning for a task that has inherently no classical analogue.

\*\*\*\*\*

Adapting Auxiliary Losses Using Gradient Similarity

Yunshu Du,Wojciech M. Czarnecki,Siddhant M. Jayakumar,Razvan Pascanu,Balaji Lakshminarayanan

One approach to deal with the statistical inefficiency of neural networks is to rely on auxiliary losses that help to build useful representations. However, it is not always trivial to know if an auxiliary task will be helpful for the main task and when it could start hurting. We propose to use the cosine similarity between gradients of tasks as an adaptive weight to detect when an auxiliary loss is helpful to the main loss. We show that our approach is guaranteed to converge to critical points of the main task and demonstrate the practical usefulness of the proposed algorithm in a few domains: multi-task supervised learning on subsets of ImageNet, reinforcement learning on gridworld, and reinforcement learning on Atari games.

\*\*\*\*\*

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  
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 inc

cluding 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.

\*\*\*\*\*

#### Hierarchically-Structured Variational Autoencoders for Long Text Generation

Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Lawrence Carin

Variational autoencoders (VAEs) have received much attention recently as an end-to-end architecture for text generation. Existing methods primarily focus on synthesizing relatively short sentences (with less than twenty words). In this paper, we propose a novel framework, hierarchically-structured variational autoencoder (hier-VAE), for generating long and coherent units of text. To enhance the model's plan-ahead ability, intermediate sentence representations are introduced into the generative networks to guide the word-level predictions. To alleviate the typical optimization challenges associated with textual VAEs, we further employ a hierarchy of stochastic layers between the encoder and decoder networks. Extensive experiments are conducted to evaluate the proposed method, where hier-VAE is shown to make effective use of the latent codes and achieve lower perplexity relative to language models. Moreover, the generated samples from hier-VAE also exhibit superior quality according to both automatic and human evaluations.

\*\*\*\*\*

#### Contextual Recurrent Convolutional Model for Robust Visual Learning

Siming Yan\*, Bowen Xiao\*, Yimeng Zhang, Tai Sing Lee

Feedforward convolutional neural network has achieved a great success in many computer vision tasks. While it validly imitates the hierarchical structure of biological visual system, it still lacks one essential architectural feature: contextual recurrent connections with feedback, which widely exists in biological visual system. In this work, we designed a Contextual Recurrent Convolutional Network with this feature embedded in a standard CNN structure. We found that such feedback connections could enable lower layers to "rethink" about their representations given the top-down contextual information. We carefully studied the components of this network, and showed its robustness and superiority over feedforward baselines in such tasks as noise image classification, partially occluded object recognition and fine-grained image classification. We believed this work could be an important step to help bridge the gap between computer vision models and real biological visual system.

\*\*\*\*\*

#### Recurrent Kalman Networks: Factorized Inference in High-Dimensional Deep Feature Spaces

Philipp Becker, Harit Pandya, Gregor H.W. Gebhardt, Cheng Zhao, Gerhard Neumann

In order to integrate uncertainty estimates into deep time-series modelling, Kalman Filters (KFs) (Kalman et al., 1960) have been integrated with deep learning models. Yet, such approaches typically rely on approximate inference techniques such as variational inference which makes learning more complex and often less scalable due to approximation errors. We propose a new deep approach to Kalman filtering which can be learned directly in an end-to-end manner using backpropagation without additional approximations. Our approach uses a high-dimensional factorized latent state representation for which the Kalman updates simplify to scalar operations and thus avoids hard to backpropagate, computationally heavy and potentially unstable matrix inversions. Moreover, we use locally linear dynamic models to efficiently propagate the latent state to the next time step. While our locally linear modelling and factorization assumptions are in general not true for the original low-dimensional state space of the system, the network finds a high-dimensional latent space where these assumptions hold to perform efficient inference. This state representation is learned jointly with the transition and noise models. The resulting network architecture, which we call Recurrent Kalman Network (RKN), can be used for any time-series data, similar to

a LSTM (Hochreiter and Schmidhuber, 1997) but uses an explicit representation of uncertainty. As shown by our experiments, the RKN obtains much more accurate uncertainty estimates than an LSTM or Gated Recurrent Units (GRUs) (Cho et al., 2014) while also showing a slightly improved prediction performance and outperform various recent generative models on an image imputation task.

\*\*\*\*\*

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

Kendall Lowrey, Aravind Rajeswaran, Sham Kakade, Emanuel Todorov, Igor Mordatch

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

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 multiplication, 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

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  $\delta$ -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 LMB and modeling images: learning representations, improving sample quality, and achieving state of the art log-likelihood on CIFAR-10 and ImageNet  $32 \times 32$ .

\*\*\*\*\*

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

Anusha Nagabandi, Chelsea Finn, Sergey Levine

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.

\*\*\*\*\*

Analysis of Memory Organization for Dynamic Neural Networks

Ying Ma, Jose Principe

An increasing number of neural memory networks have been developed, leading to the need for a systematic approach to analyze and compare their underlying memory capabilities. Thus, in this paper, we propose a taxonomy for four popular dynamic models: vanilla recurrent neural network, long short-term memory, neural stack and neural RAM and their variants. Based on this taxonomy, we create a framework to analyze memory organization and then compare these network architectures. This analysis elucidates how different mapping functions capture the information in the past of the input, and helps to open the dynamic neural network black box from the perspective of memory usage. Four representative tasks that would fit optimally the characteristics of each memory network are carefully selected to show each network's expressive power. We also discuss how to use this taxonomy to help users select the most parsimonious type of memory network for a specific task. Two natural language processing applications are used to evaluate the methodology in a realistic setting.

\*\*\*\*\*

Expanding the Reach of Federated Learning by Reducing Client Resource Requirements

Sebastian Caldas, Jakub Konečný, Brendan McMahan, Ameet Talwalkar

Communication on heterogeneous edge networks is a fundamental bottleneck in Federated Learning (FL), restricting both model capacity and user participation. To address this issue, we introduce two novel strategies to reduce communication costs: (1) the use of lossy compression on the global model sent server-to-client; and (2) Federated Dropout, which allows users to efficiently train locally on smaller subsets of the global model and also provides a reduction in both client-to-server communication and local computation. We empirically show that these strategies, combined with existing compression approaches for client-to-server communication, collectively provide up to a 9.6x reduction in server-to-client communication, a 1.5x reduction in local computation, and a 24x reduction in upload communication, all without degrading the quality of the final model. We thus comprehensively reduce FL's impact on client device resources, allowing higher capacity models to be trained, and a more diverse set of users to be reached.

\*\*\*\*\*

DEEP GRAPH TRANSLATION

Xiaojie Guo, Lingfei Wu, Liang Zhao

The tremendous success of deep generative models on generating continuous data

like image and audio has been achieved; however, few deep graph generative models have been proposed to generate discrete data such as graphs. The recently proposed approaches are typically unconditioned generative models which have no control over modes of the graphs being generated. Differently, in this paper, we are interested in a new problem named Deep Graph Translation: given an input graph, the goal is to infer a target graph by learning their underlying translation mapping. Graph translation could be highly desirable in many applications such as disaster management and rare event forecasting, where the rare and abnormal graph patterns (e.g., traffic congestions and terrorism events) will be inferred prior to their occurrence even without historical data on the abnormal patterns for this specific graph (e.g., a road network or human contact network). To this end, we propose a novel Graph-Translation-Generative Adversarial Networks (GT-GAN) which translates one mode of the input graphs to its target mode. GT-GAN consists of a graph translator where we propose new graph convolution and deconvolution layers to learn the global and local translation mapping. A new conditional graph discriminator has also been proposed to classify target graphs by conditioning on input graphs. Extensive experiments on multiple synthetic and real-world datasets demonstrate the effectiveness and scalability of the proposed GT-GAN.

\*\*\*\*\*

#### Value Propagation Networks

Nantas Nardelli, Gabriel Synnaeve, Zeming Lin, Pushmeet Kohli, Philip H. S. Torr, Nicolas Usunier

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.

\*\*\*\*\*

#### Online Hyperparameter Adaptation via Amortized Proximal Optimization

Paul Vicol, Jeffery Z. HaoChen, Roger Grosse

Effective performance of neural networks depends critically on effective tuning of optimization hyperparameters, especially learning rates (and schedules thereof). We present Amortized Proximal Optimization (APO), which takes the perspective that each optimization step should approximately minimize a proximal objective (similar to the ones used to motivate natural gradient and trust region policy optimization). Optimization hyperparameters are adapted to best minimize the proximal objective after one weight update. We show that an idealized version of APO (where an oracle minimizes the proximal objective exactly) achieves global convergence to stationary point and locally second-order convergence to global optimum for neural networks. APO incurs minimal computational overhead. We experiment with using APO to adapt a variety of optimization hyperparameters online during training, including (possibly layer-specific) learning rates, damping coefficients, and gradient variance exponents. For a variety of network architectures and optimization algorithms (including SGD, RMSprop, and K-FAC), we show that with minimal tuning, APO performs competitively with carefully tuned optimizers.

\*\*\*\*\*

#### Efficient Augmentation via Data Subsampling

Michael Kuchnik, Virginia Smith

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.

\*\*\*\*\*

Complementary-label learning for arbitrary losses and models

Takashi Ishida, Gang Niu, Aditya Krishna Menon, Masashi Sugiyama

In contrast to the standard classification paradigm where the true (or possibly noisy) class is given to each training pattern, complementary-label learning only uses training patterns each equipped with a complementary label. This only specifies one of the classes that the pattern does not belong to. The seminal paper on complementary-label learning proposed an unbiased estimator of the classification risk that can be computed only from complementarily labeled data. However, it required a restrictive condition on the loss functions, making it impossible to use popular losses such as the softmax cross-entropy loss. Recently, another formulation with the softmax cross-entropy loss was proposed with consistency guarantee. However, this formulation does not explicitly involve a risk estimator. Thus model/hyper-parameter selection is not possible by cross-validation—we may need additional ordinarily labeled data for validation purposes, which is not available in the current setup. In this paper, we give a novel general framework of complementary-label learning, and derive an unbiased risk estimator for arbitrary losses and models. We further improve the risk estimator by non-negative correction and demonstrate its superiority through experiments.

\*\*\*\*\*

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

Jialin Liu, Xiaohan Chen, Zhangyang Wang, Wotao Yin

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.

\*\*\*\*\*

Language Model Pre-training for Hierarchical Document Representations

Ming-Wei Chang, Kristina Toutanova, Kenton Lee, Jacob Devlin

Hierarchical neural architectures can efficiently capture long-distance dependencies and have been used for many document-level tasks such as summarization, document segmentation, and fine-grained sentiment analysis. However, effective usage of such a large context can be difficult to learn, especially in the case where there is limited labeled data available.

Building on the recent success of language model pretraining methods for learning flat representations of text, we propose algorithms for pre-training hierarchical document representations from unlabeled data. Unlike prior work, which has focused on pre-training contextual token representations or context-independent sentence/paragraph representations, our hierarchical document representations inc

lude fixed-length sentence/paragraph representations which integrate contextual information from the entire documents. Experiments on document segmentation, document-level question answering, and extractive document summarization demonstrate the effectiveness of the proposed pre-training algorithms.

\*\*\*\*\*

#### Non-Synergistic Variational Autoencoders

Gonzalo Barrientos, Sten Sootla

Learning disentangling representations of the independent factors of variations that explain the data in an unsupervised setting is still a major challenge. In the following paper we address the task of disentanglement and introduce a new state-of-the-art approach called Non-synergistic variational Autoencoder (Non-Syn VAE). Our model draws inspiration from population coding, where the notion of synergy arises when we describe the encoded information by neurons in the form of responses from the stimuli. If those responses convey more information together than separate as independent sources of encoding information, they are acting synergistically. By penalizing the synergistic mutual information within the latents we encourage information independence and by doing that disentangle the latent factors. Notably, our approach could be added to the VAE framework easily, where the new ELBO function is still a lower bound on the log likelihood. In addition, we qualitatively compare our model with Factor VAE and show that this one implicitly minimises the synergy of the latents.

\*\*\*\*\*

#### Approximation and non-parametric estimation of ResNet-type convolutional neural networks via block-sparse fully-connected neural networks

Kenta Oono, Taiji Suzuki

We develop new approximation and statistical learning theories of convolutional neural networks (CNNs) via the ResNet-type structure where the channel size, filter size, and width are fixed. It is shown that a ResNet-type CNN is a universal approximator and its expression ability is no worse than fully-connected neural networks (FNNs) with a  $\text{block-sparse}$  structure even if the size of each layer in the CNN is fixed. Our result is general in the sense that we can automatically translate any approximation rate achieved by block-sparse FNNs into that by CNNs. Thanks to the general theory, it is shown that learning on CNNs satisfies optimality in approximation and estimation of several important function classes.

As applications, we consider two types of function classes to be estimated: the Barron class and  $H^{\text{older}}$  class. We prove the clipped empirical risk minimization (ERM) estimator can achieve the same rate as FNNs even the channel size, filter size, and width of CNNs are constant with respect to the sample size. This is minimax optimal (up to logarithmic factors) for the  $H^{\text{older}}$  class. Our proof is based on sophisticated evaluations of the covering number of CNNs and the non-trivial parameter rescaling technique to control the Lipschitz constant of CNNs to be constructed.

\*\*\*\*\*

#### Deep Imitative Models for Flexible Inference, Planning, and Control

Nicholas Rhinehart, Rowan McAllister, Sergey Levine

Imitation learning provides an appealing framework for autonomous control: in many tasks, demonstrations of preferred behavior can be readily obtained from human experts, removing the need for costly and potentially dangerous online data collection in the real world. However, policies learned with imitation learning have limited flexibility to accommodate varied goals at test time. Model-based reinforcement learning (MBRL) offers considerably more flexibility, since a predictive model learned from data can be used to achieve various goals at test time. However, MBRL suffers from two shortcomings. First, the model does not help to choose desired or safe outcomes -- its dynamics estimate only what is possible, not what is preferred. Second, MBRL typically requires additional online data collection to ensure that the model is accurate in those situations that are actually encountered when attempting to achieve test time goals. Collecting this data with a partially trained model can be dangerous and time-consuming. In this paper



, we aim to combine the benefits of imitation learning and MBRL, and propose imitative models: probabilistic predictive models able to plan expert-like trajectories to achieve arbitrary goals. We find this method substantially outperforms both direct imitation and MBRL in a simulated autonomous driving task, and can be learned efficiently from a fixed set of expert demonstrations without additional online data collection. We also show our model can flexibly incorporate user-supplied costs at test-time, can plan to sequences of goals, and can even perform well with imprecise goals, including goals on the wrong side of the road.

\*\*\*\*\*

Recall Traces: Backtracking Models for Efficient Reinforcement Learning

Anirudh Goyal, Philemon Brakel, William Fedus, Soumye Singh, Timothy Lillicrap, Sergey Levine, Hugo Larochelle, Yoshua Bengio

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 `\texttt{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.

\*\*\*\*\*

Learning Goal-Conditioned Value Functions with one-step Path rewards rather than Goal-Rewards

Vikas Dhiman, Shurjo Banerjee, Jeffrey M Siskind, Jason J Corso

Multi-goal reinforcement learning (MGRL) addresses tasks where the desired goal state can change for every trial. State-of-the-art algorithms model these problems such that the reward formulation depends on the goals, to associate them with high reward. This dependence introduces additional goal reward resampling steps in algorithms like Hindsight Experience Replay (HER) that reuse trials in which the agent fails to reach the goal by recomputing rewards as if reached states were pseudo-desired goals. We propose a reformulation of goal-conditioned value functions for MGRL that yields a similar algorithm, while removing the dependence of reward functions on the goal. Our formulation thus obviates the requirement of reward-recomputation that is needed by HER and its extensions. We also extend a closely related algorithm, Floyd-Warshall Reinforcement Learning, from tabular domains to deep neural networks for use as a baseline. Our results are competitive with HER while substantially improving sampling efficiency in terms of reward computation.

\*\*\*\*\*

Fixup Initialization: Residual Learning Without Normalization

Hongyi Zhang, Yann N. Dauphin, Tengyu Ma

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 p

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

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.

\*\*\*\*\*

#### A Self-Supervised Method for Mapping Human Instructions to Robot Policies

Hsin-Wei Yu, Po-Yu Wu, Chih-An Tsao, You-An Shen, Shih-Hsuan Lin, Zhang-Wei Hong, Yi-Hsiang Chang, Chun-Yi Lee

In this paper, we propose a modular approach which separates the instruction-to-action mapping procedure into two separate stages. The two stages are bridged via an intermediate representation called a goal, which stands for the result after a robot performs a specific task.

The first stage maps an input instruction to a goal, while the second stage maps the goal to an appropriate policy selected from a set of robot policies. The policy is selected with an aim to guide the robot to reach the goal as close as possible. We implement the above two stages as a framework consisting of two distinct modules: an instruction-goal mapping module and a goal-policy mapping module. Given a human instruction in the evaluation phase, the instruction-goal mapping module first translates the instruction to a robot-interpretable goal. Once a goal is derived by the instruction-goal mapping module, the goal-policy mapping module then follows up to search through the goal-policy pairs to look for policy to be mapped by the instruction. Our experimental results show that the proposed method is able to learn an effective instruction-to-action mapping procedure in an environment with a given instruction set more efficiently than the baselines. In addition to the impressive data-efficiency, the results also show that our method can be adapted to a new instruction set and a new robot actionspace much faster than the baselines. The evidence suggests that our modular approach does lead to better adaptability and efficiency.

\*\*\*\*\*

#### 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

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> ).

\*\*\*\*\*

#### Transferring SLU Models in Novel Domains

Yaohua Tang,Kaixiang Mo,Qian Xu,Chao Zhang,Qiang Yang

Spoken language understanding (SLU) is a critical component in building dialogue systems. When building models for novel natural language domains, a major challenge is the lack of data in the new domains, no matter whether the data is annotated or not. Recognizing and annotating ``intent'' and ``slot'' of natural languages is a time-consuming process. Therefore, spoken language understanding in low resource domains remains a crucial problem to address. In this paper, we address this problem by proposing a transfer-learning method, whereby a SLU model is transferred to a novel but data-poor domain via a deep neural network framework.

We also introduce meta-learning in our work to bridge the semantic relations between seen and unseen data, allowing new intents to be recognized and new slots to be filled with much lower new training effort. We show the performance improvement with extensive experimental results for spoken language understanding in low resource domains. We show that our method can also handle novel intent recognition and slot-filling tasks. Our methodology provides a feasible solution for alleviating data shortages in spoken language understanding.

\*\*\*\*\*

#### End-to-End Multi-Lingual Multi-Speaker Speech Recognition

Hiroshi Seki,Takaaki Hori,Shinji Watanabe,Jonathan Le Roux,John R. Hershey

The expressive power of end-to-end automatic speech recognition (ASR) systems enables direct estimation of the character or word label sequence from a sequence of acoustic features. Direct optimization of the whole system is advantageous because it not only eliminates the internal linkage necessary for hybrid systems,

but also extends the scope of potential application use cases by training the model for multiple objectives. Several multi-lingual ASR systems were recently proposed based on a monolithic neural network architecture without language-dependent modules, showing that modeling of multiple languages is well within the capabilities of an end-to-end framework. There has also been growing interest in multi-speaker speech recognition, which enables generation of multiple label sequences from single-channel mixed speech. In particular, a multi-speaker end-to-end ASR system that can directly model one-to-many mappings without additional auxiliary clues was recently proposed. In this paper, we propose an all-in-one end-to-end multi-lingual multi-speaker ASR system that integrates the capabilities of these two systems. The proposed model is evaluated using mixtures of two speakers generated by using 10 languages, including mixed-language utterances.

\*\*\*\*\*

#### W2GAN: RECOVERING AN OPTIMAL TRANSPORT MAP WITH A GAN

Leygonie Jacob\*,Jennifer She\*,Amjad Almahairi,Sai Rajeswar,Aaron Courville

Understanding and improving Generative Adversarial Networks (GAN) using notions from Optimal Transport (OT) theory has been a successful area of study, originally established by the introduction of the Wasserstein GAN (WGAN). An increasing number of GANs incorporate OT for improving their discriminators, but that is so far the sole way for the two domains to cross-fertilize. In this work we address the converse question: is it possible to recover an optimal map in a GAN fashion? To achieve this, we build a new model relying on the second Wasserstein distance. This choice enables the use of many results from OT community. In particular, we may completely describe the dynamics of the generator during training. In addition, experiments show that practical uses of our model abide by the rule of evolution we describe. As an application, our generator may be considered as a new way of computing an optimal transport map. It is competitive in low-dimension with standard and deterministic ways to approach the same problem. In high dimension, the fact it is a GAN-style method makes it more powerful than other methods.

\*\*\*\*\*

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

Hsin-Yuan Huang, Eunsol Choi, Wen-tau Yih

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.

\*\*\*\*\*

## Language Modeling with Graph Temporal Convolutional Networks

Hongyin Luo, Yichen Li, Jie Fu, James Glass

Recently, there have been some attempts to use non-recurrent neural models for language modeling.

However, a noticeable performance gap still remains.

We propose a non-recurrent neural language model, dubbed graph temporal convolutional network (GTCN), that relies on graph neural network blocks and convolution operations. While the standard recurrent neural network language models encode sentences sequentially without modeling higher-level structural information, our model regards sentences as graphs and processes input words within a message propagation framework, aiming to learn better syntactic information by inferring skip-word connections. Specifically, the graph network blocks operate in parallel and learn the underlying graph structures in sentences without any additional annotation pertaining to structure knowledge. Experiments demonstrate that the model without recurrence can achieve comparable perplexity results in language modeling tasks and successfully learn syntactic information.

\*\*\*\*\*

## Probabilistic Planning with Sequential Monte Carlo methods

Alexandre Piche, Valentin Thomas, Cyril Ibrahim, Yoshua Bengio, Chris Pal

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.

\*\*\*\*\*

## A theoretical framework for deep and locally connected ReLU network

Yuandong Tian

Understanding theoretical properties of deep and locally connected nonlinear network, such as deep convolutional neural network (DCNN), is still a hard problem despite its empirical success. In this paper, we propose a novel theoretical framework for such networks with ReLU nonlinearity. The framework bridges data distribution with gradient descent rules, favors disentangled representations and is compatible with common regularization techniques such as Batch Norm, after a novel discovery of its projection nature. The framework is built upon teacher-student setting, by projecting the student's forward/backward pass onto the teacher's computational graph. We do not impose unrealistic assumptions (e.g., Gaussian inputs, independence of activation, etc). Our framework could help facilitate theoretical analysis of many practical issues, e.g. disentangled representations in deep networks.

\*\*\*\*\*

## Unsupervised Video-to-Video Translation

Dina Bashkirova, Ben Usman, Kate Saenko

Unsupervised image-to-image translation is a recently proposed task of translating an image to a different style or domain given only unpaired image examples at training time. In this paper, we formulate a new task of unsupervised video-to-video translation, which poses its own unique challenges. Translating video implies learning not only the appearance of objects and scenes but also realistic motion and transitions between consecutive frames. We investigate the performance of per-frame video-to-video translation using existing image-to-image translation networks, and propose a spatio-temporal 3D translator as an alternative solution to this problem. We evaluate our 3D method on multiple synthetic datasets, such as moving colorized digits, as well as the realistic segmentation-to-video GT-A dataset and a new CT-to-MRI volumetric images translation dataset. Our results show that frame-wise translation produces realistic results on a single frame level but underperforms significantly on the scale of the whole video compared to our three-dimensional translation approach, which is better able to learn the complex structure of video and motion and continuity of object appearance.

\*\*\*\*\*

Spreading vectors for similarity search

Alexandre Sablayrolles,Matthijs Douze,Cordelia Schmid,Hervé Jégou

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. Furthermore, 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

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.

\*\*\*\*\*

Double Neural Counterfactual Regret Minimization

Hui Li,Kailiang Hu,Zhibang Ge,Tao Jiang,Yuan Qi,Le Song

Counterfactual regret minimization (CRF) is a fundamental and effective technique for solving imperfect information games. However, the original CRF algorithm only works for discrete state and action spaces, and the resulting strategy is maintained as a tabular representation. Such tabular representation limits the method from being directly applied to large games and continuing to improve from a poor strategy profile. In this paper, we propose a double neural representation for the Imperfect Information Games, where one neural network represents the cum

ulative regret, and the other represents the average strategy. Furthermore, we adopt the counterfactual regret minimization algorithm to optimize this double neural representation. To make neural learning efficient, we also developed several novel techniques including a robust sampling method, mini-batch Monte Carlo counterfactual regret minimization (MCCFR) and Monte Carlo counterfactual regret minimization plus (MCCFR+) which may be of independent interests. Experimentally, we demonstrate that the proposed double neural algorithm converges significantly better than the reinforcement learning counterpart.

\*\*\*\*\*

Learning powerful policies and better dynamics models by encouraging consistency  
Shagun Sodhani, Anirudh Goyal, Tristan Deleu, Yoshua Bengio, Jian Tang

Model-based reinforcement learning approaches have the promise of being sample efficient. Much of the progress in learning dynamics models in RL has been made by learning models via supervised learning. There is enough evidence that humans build a model of the environment, not only by observing the environment but also by interacting with the environment. Interaction with the environment allows humans to carry out "experiments": taking actions that help uncover true causal relationships which can be used for building better dynamics models. Analogously, we would expect such interaction to be helpful for a learning agent while learning to model the environment dynamics. In this paper, we build upon this intuition, by using an auxiliary cost function to ensure consistency between what the agent observes (by acting in the real world) and what it imagines (by acting in the "learned" world). Our empirical analysis shows that the proposed approach helps to train powerful policies as well as better dynamics models.

\*\*\*\*\*

Human-Guided Column Networks: Augmenting Deep Learning with Advice

Mayukh Das, Yang Yu, Devendra Singh Dhami, Gautam Kunapuli, Sriraam Natarajan

While extremely successful in several applications, especially with low-level representations; sparse, noisy samples and structured domains (with multiple objects and interactions) are some of the open challenges in most deep models. Column Networks, a deep architecture, can succinctly capture such domain structure and interactions, but may still be prone to sub-optimal learning from sparse and noisy samples. Inspired by the success of human-advice guided learning in AI, especially in data-scarce domains, we propose Knowledge-augmented Column Networks that leverage human advice/knowledge for better learning with noisy/sparse samples. Our experiments demonstrate how our approach leads to either superior overall performance or faster convergence.

\*\*\*\*\*

Learning what you can do before doing anything

Oleh Rybkin, Karl Pertsch, Konstantinos G. Derpanis, Kostas Daniilidis, Andrew Jaegle

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](https://daniilidis-group.github.io/learned_action_spaces)

\*\*\*\*\*

A Frank-Wolfe Framework for Efficient and Effective Adversarial Attacks

Jinghui Chen, Jinfeng Yi, Quanquan Gu

Depending on how much information an adversary can access to, adversarial attacks can be classified as white-box attack and black-box attack. In both cases, optimization-based attack algorithms can achieve relatively low distortions and high attack success rates. However, they usually suffer from poor time and query complexities, thereby limiting their practical usefulness. In this work, we focus on the problem of developing efficient and effective optimization-based adversarial attack algorithms. In particular, we propose a novel adversarial attack framework for both white-box and black-box settings based on the non-convex Frank-Wolfe algorithm. We show in theory that the proposed attack algorithms are efficient with an  $O(1/\sqrt{T})$  convergence rate. The empirical results of attacking Inception V3 model and ResNet V2 model on the ImageNet dataset also verify the efficiency and effectiveness of the proposed algorithms. More specific, our proposed algorithms attain the highest attack success rate in both white-box and black-box attacks among all baselines, and are more time and query efficient than the state-of-the-art.

\*\*\*\*\*

Lagging Inference Networks and Posterior Collapse in Variational Autoencoders

Junxian He, Daniel Spokoyny, Graham Neubig, Taylor Berg-Kirkpatrick

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 Krashenninikov, Jordan Alexander, Pieter Abbeel, Anca Dragan

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>.

\*\*\*\*\*

On Inductive Biases in Deep Reinforcement Learning

Matteo Hessel, Hado van Hasselt, Joseph Modayil, David Silver

Many deep reinforcement learning algorithms contain inductive biases that sculpt

the agent's objective and its interface to the environment. These inductive biases can take many forms, including domain knowledge and pretuned hyper-parameters. In general, there is a trade-off between generality and performance when we use such biases. Stronger biases can lead to faster learning, but weaker biases can potentially lead to more general algorithms that work on a wider class of problems.

This trade-off is relevant because these inductive biases are not free; substantial effort may be required to obtain relevant domain knowledge or to tune hyper-parameters effectively. In this paper, we re-examine several domain-specific components that modify the agent's objective and environmental interface. We investigated whether the performance deteriorates when all these fixed components are replaced with adaptive solutions from the literature. In our experiments, performance sometimes decreased with the adaptive components, as one might expect when comparing to components crafted for the domain, but sometimes the adaptive components performed better. We then investigated the main benefit of having fewer domain-specific components, by comparing the learning performance of the two systems on a different set of continuous control problems, without additional tuning of either system. As hypothesized, the system with adaptive components performed better on many of the tasks.

\*\*\*\*\*

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

Huaxia Wang, Chun-Nam Yu

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.

\*\*\*\*\*

#### Learning Abstract Models for Long-Horizon Exploration

Evan Zheran Liu, Ramtin Keramati, Sudarshan Seshadri, Kelvin Guu, Panupong Pasupat, Emma Brunskill, Percy Liang

In high-dimensional reinforcement learning settings with sparse rewards, performing effective exploration to even obtain any reward signal is an open challenge. While model-based approaches hold promise of better exploration via planning, it is extremely difficult to learn a reliable enough Markov Decision Process (MDP) in high dimensions (e.g., over  $10^{100}$  states). In this paper, we propose learning an abstract MDP over a much smaller number of states (e.g.,  $10^5$ ), which we can plan over for effective exploration. We assume we have an abstraction function that maps concrete states (e.g., raw pixels) to abstract states (e.g., agent position, ignoring other objects). In our approach, a manager maintains an abstract MDP over a subset of the abstract states, which grows monotonically through targeted exploration (possible due to the abstract MDP). Concurrently, we learn a worker policy to travel between abstract states; the worker deals with the messiness of concrete states and presents a clean abstraction to the manager. On three of the hardest games from the Arcade Learning Environment (Montezuma's, Pitfall!, and Private Eye), our approach outperforms the previous



state-of-the-art by over a factor of 2 in each game. In Pitfall!, our approach is the first to achieve superhuman performance without demonstrations.

\*\*\*\*\*

Learning to remember: Dynamic Generative Memory for Continual Learning  
Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Moin Nabi

Continuously trainable models should be able to learn from a stream of data over an undefined period of time. This becomes even more difficult in a strictly incremental context, where data access to previously seen categories is not possible. To that end, we propose making use of a conditional generative adversarial model where the generator is used as a memory module through neural masking to emulate neural plasticity in the human brain. This memory module is further associated with a dynamic capacity expansion mechanism. Taken together, this method facilitates a resource efficient capacity adaption to accommodate new tasks, while retaining previously attained knowledge. The proposed approach outperforms state-of-the-art algorithms on publicly available datasets, overcoming catastrophic forgetting.

\*\*\*\*\*

Improving the Generalization of Adversarial Training with Domain Adaptation  
Chuanbiao Song, Kun He, Liwei Wang, John E. Hopcroft

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 by 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-Adversarial Training (PAT) and the defense performance is improved considerably.

\*\*\*\*\*

NSGA-Net: A Multi-Objective Genetic Algorithm for Neural Architecture Search  
Zhichao Lu, Ian Whalen, Vishnu Boddeti, Yashesh Dhebar, Kalyanmoy Deb, Erik Goodman, Wolfgang Banzhaf

This paper introduces NSGA-Net, an evolutionary approach for neural architecture search (NAS). NSGA-Net is designed with three goals in mind: (1) a NAS procedure for multiple, possibly conflicting, objectives, (2) efficient exploration and exploitation of the space of potential neural network architectures, and (3) output of a diverse set of network architectures spanning a trade-off frontier of the objectives in a single run. NSGA-Net is a population-based search algorithm that explores a space of potential neural network architectures in three steps, namely, a population initialization step that is based on prior-knowledge from hand-crafted architectures, an exploration step comprising crossover and mutation of architectures and finally an exploitation step that applies the entire history of evaluated neural architectures in the form of a Bayesian Network prior. Experimental results suggest that combining the objectives of minimizing both an error metric and computational complexity, as measured by FLOPS, allows NSGA-Net to find competitive neural architectures near the Pareto front of both objectives

on two different tasks, object classification and object alignment. NSGA-Net obtains networks that achieve 3.72% (at 4.5 million FLOP) error on CIFAR-10 classification and 8.64% (at 26.6 million FLOP) error on the CMU-Car alignment task.

\*\*\*\*\*

The Expressive Power of Deep Neural Networks with Circulant Matrices

Alexandre Araujo, Benjamin Negrevergne, Yann Chevalere, Jamal Atif

Recent results from linear algebra stating that any matrix can be decomposed into products of diagonal and circulant matrices has led to the design of compact deep neural network architectures that perform well in practice. In this paper, we bridge the gap between these good empirical results

and the theoretical approximation capabilities of Deep diagonal-circulant ReLU networks. More precisely, we first demonstrate that a Deep diagonal-circulant ReLU networks of

bounded width and small depth can approximate a deep ReLU network in which the dense matrices are

of low rank. Based on this result, we provide new bounds on the expressive power and universal approximateness of this type of networks. We support our experimental results with thorough experiments on a large, real world video classification problem.

\*\*\*\*\*

Uncertainty-guided Lifelong Learning in Bayesian Networks

Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, Marcus Rohrbach

Sequentially learning of tasks arriving in a continuous stream is a complex problem and becomes more challenging when the model has a fixed capacity. Lifelong learning aims at learning new tasks without forgetting previously learnt ones as well as freeing up capacity for learning future tasks. We argue that identifying the most influential parameters in a representation learned for one task plays a critical role to decide on what to remember for continual learning.

Motivated by the statistically-grounded uncertainty defined in Bayesian neural networks, we propose to formulate a Bayesian lifelong learning framework,  $\text{BLLL}$ , that addresses two lifelong learning directions: 1) completely eliminating catastrophic forgetting using weight pruning, where a hard selection mask freezes the most certain parameters ( $\text{BLLL-PRN}$ ) and 2) reducing catastrophic forgetting by adaptively regularizing the learning rates using the parameter uncertainty ( $\text{BLLL-REG}$ ). While  $\text{BLLL-PRN}$  is by definition a zero-forgetting guaranteed method,  $\text{BLLL-REG}$ , despite exhibiting some small forgetting, is a task-agnostic lifelong learner, which does not require to know when a new task arrives. This feature makes  $\text{BLLL-REG}$  a more convenient candidate for applications such as robotics or on-line learning in which such information is not available. We evaluate our Bayesian learning approaches extensively on diverse object classification datasets in short and long sequences of tasks and perform superior or marginally better than the existing approaches.

\*\*\*\*\*

Importance Resampling for Off-policy Policy Evaluation

Matthew Schlegel, Wesley Chung, Daniel Graves, Martha White

Importance sampling is a common approach to off-policy learning in reinforcement learning. While it is consistent and unbiased, it can result in high variance updates to the parameters for the value function. Weighted importance sampling (WIS) has been explored to reduce variance for off-policy policy evaluation, but only for linear value function approximation. In this work, we explore a resampling strategy to reduce variance, rather than a reweighting strategy. We propose Importance Resampling (IR) for off-policy learning, that resamples experience from the replay buffer and applies a standard on-policy update. The approach avoids using importance sampling ratios directly in the update, instead correcting the distribution over transitions before the update. We characterize the bias and consistency of our estimator, particularly compared to WIS. We then demonstrate in several toy domains that IR has improved sample efficiency and parameter sensitivity, as compared to several baseline WIS estimators and to IS. We conclude with a demonstration showing IR improves over IS for learning a value function from images in a racing car simulator.

\*\*\*\*\*

#### Probabilistic Federated Neural Matching

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, Yasin Saman Khazaeni

In federated learning problems, data is scattered across different servers and exchanging or pooling it is often impractical or prohibited. We develop a Bayesian nonparametric framework for federated learning with neural networks. Each data server is assumed to train local neural network weights, which are modeled through our framework. We then develop an inference approach that allows us to synthesize a more expressive global network without additional supervision or data pooling. We then demonstrate the efficacy of our approach on federated learning problems simulated from two popular image classification datasets.

\*\*\*\*\*

#### Evolutionary-Neural Hybrid Agents for Architecture Search

Krzysztof Maziarczyk, Andrey Khorlin, Quentin de Laroussilhe, Andrea Gesmundo

Neural Architecture Search has recently shown potential to automate the design of Neural Networks. The use of Neural Network agents trained with Reinforcement Learning can offer the possibility to learn complex patterns, as well as the ability to explore a vast and compositional search space. On the other hand, evolutionary algorithms offer the greediness and sample efficiency needed for such an application, as each sample requires a considerable amount of resources. We propose a class of Evolutionary-Neural hybrid agents (Evo-NAS), that retain the best qualities of the two approaches. We show that the Evo-NAS agent can outperform both Neural and Evolutionary agents, both on a synthetic task, and on architecture search for a suite of text classification datasets.

\*\*\*\*\*

#### CAML: Fast Context Adaptation via Meta-Learning

Luisa M Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, Shimon Whiteson

We propose CAML, a meta-learning method for fast adaptation that partitions the model parameters into two parts: context parameters that serve as additional input to the model and are adapted on individual tasks, and shared parameters that are meta-trained and shared across tasks. At test time, the context parameters are updated with one or several gradient steps on a task-specific loss that is backpropagated through the shared part of the network. Compared to approaches that adjust all parameters on a new task (e.g., MAML), our method can be scaled up to larger networks without overfitting on a single task, is easier to implement, and saves memory writes during training and network communication at test time for distributed machine learning systems. We show empirically that this approach outperforms MAML, is less sensitive to the task-specific learning rate, can capture meaningful task embeddings with the context parameters, and outperforms alternative partitionings of the parameter vectors.

\*\*\*\*\*

#### Subgradient Descent Learns Orthogonal Dictionaries

Yu Bai, Qijia Jiang, Ju Sun

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

The Differentiable Neural Computer (DNC) can learn algorithmic and question answer

ering 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%.

\*\*\*\*\*

Exploring the interpretability of LSTM neural networks over multi-variable data  
Tian Guo,Tao Lin

In learning a predictive model over multivariate time series consisting of target and exogenous variables, the forecasting performance and interpretability of the model are both essential for deployment and uncovering knowledge behind the data.

To this end, we propose the interpretable multi-variable LSTM recurrent neural network (IMV-LSTM) capable of providing accurate forecasting as well as both temporal and variable level importance interpretation.

In particular, IMV-LSTM is equipped with tensorized hidden states and update process, so as to learn variables-wise hidden states.

On top of it, we develop a mixture attention mechanism and associated summarization methods to quantify the temporal and variable importance in data.

Extensive experiments using real datasets demonstrate the prediction performance and interpretability of IMV-LSTM in comparison to a variety of baselines.

It also exhibits the prospect as an end-to-end framework for both forecasting and knowledge extraction over multi-variate data.

\*\*\*\*\*

Unsupervised Control Through Non-Parametric Discriminative Rewards

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

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.

\*\*\*\*\*

Surprising Negative Results for Generative Adversarial Tree Search

Kamyar Azizzadenesheli,Brandon Yang,Weitang Liu,Emma Brunskill,Zachary Lipton,Anima Anandkumar

While many recent advances in deep reinforcement learning rely on model-free methods, model-based approaches remain an alluring prospect for their potential to exploit unsupervised data to learn environment dynamics. One prospect is to pursue hybrid approaches, as in AlphaGo, which combines Monte-Carlo Tree Search (MCTS)—a model-based method—with deep-Q networks (DQNs)—a model-free method. MCTS requires generating rollouts, which is computationally expensive. In this paper, we propose to simulate roll-outs, exploiting the latest breakthroughs in image-to-image transduction, namely Pix2Pix GANs, to predict the dynamics of the environment. Our proposed algorithm, generative adversarial tree search (GATS), simulates rollouts up to a specified depth using both a GAN-based dynamics model and a reward predictor. GATS employs MCTS for planning over the simulated samples and uses DQN to estimate the Q-function at the leaf states. Our theoretical analysis establishes some favorable properties of GATS vis-a-vis the bias-variance trade-off and empirical results show that on 5 popular Atari games, the dynamics and

reward predictors converge quickly to accurate solutions. However, GATS fails to outperform DQNs in 4 out of 5 games. Notably, in these experiments, MCTS has only short rollouts (up to tree depth 4), while previous successes of MCTS have involved tree depth in the hundreds. We present a hypothesis for why tree search with short rollouts can fail even given perfect modeling.

\*\*\*\*\*

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

Matthew Mackay, Paul Vicol, Jonathan Lorraine, David Duvenaud, Roger Grosse

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).

\*\*\*\*\*

#### Optimal Attacks against Multiple Classifiers

Juan C. Perdomo, Yaron Singer

We study the problem of designing provably optimal adversarial noise algorithms that induce misclassification in settings where a learner aggregates decisions from multiple classifiers. Given the demonstrated vulnerability of state-of-the-art models to adversarial examples, recent efforts within the field of robust machine learning have focused on the use of ensemble classifiers as a way of boosting the robustness of individual models. In this paper, we design provably optimal attacks against a set of classifiers. We demonstrate how this problem can be framed as finding strategies at equilibrium in a two player, zero sum game between a learner and an adversary and consequently illustrate the need for randomization in adversarial attacks. The main technical challenge we consider is the design of best response oracles that can be implemented in a Multiplicative Weight Updates framework to find equilibrium strategies in the zero-sum game. We develop a series of scalable noise generation algorithms for deep neural networks, and show that it outperforms state-of-the-art attacks on various image classification tasks. Although there are generally no guarantees for deep learning, we show this is a well-principled approach in that it is provably optimal for linear classifiers. The main insight is a geometric characterization of the decision space that reduces the problem of designing best response oracles to minimizing a quadratic function over a set of convex polytopes.

\*\*\*\*\*

#### Classification in the dark using tactile exploration

Mayur Mudigonda, Blake Tickell, Pulkit Agrawal

Combining information from different sensory modalities to execute goal directed actions is a key aspect of human intelligence. Specifically, human agents are very easily able to translate the task communicated in one sensory domain (say vision) into a representation that enables them to complete this task when they can only sense their environment using a separate sensory modality (say touch). In order to build agents with similar capabilities, in this work we consider the p

problem of a retrieving a target object from a drawer. The agent is provided with an image of a previously unseen object and it explores objects in the drawer using only tactile sensing to retrieve the object that was shown in the image without receiving any visual feedback. Success at this task requires close integration of visual and tactile sensing. We present a method for performing this task in a simulated environment using an anthropomorphic hand. We hope that future research in the direction of combining sensory signals for acting will find the object retrieval from a drawer to be a useful benchmark problem

\*\*\*\*\*

Scalable Neural Theorem Proving on Knowledge Bases and Natural Language  
Pasquale Minervini, Matko Bosnjak, Tim Rocktäschel, Edward Grefenstette, Sebastian Riedel

Reasoning over text and Knowledge Bases (KBs) is a major challenge for Artificial Intelligence, with applications in machine reading, dialogue, and question answering. Transducing text to logical forms which can be operated on is a brittle and error-prone process. Operating directly on text by jointly learning representations and transformations thereof by means of neural architectures that lack the ability to learn and exploit general rules can be very data-inefficient and not generalise correctly. These issues are addressed by Neural Theorem Provers (NTPs) (Rocktäschel & Riedel, 2017), neuro-symbolic systems based on a continuous relaxation of Prolog’s backward chaining algorithm, where symbolic unification between atoms is replaced by a differentiable operator computing the similarity between their embedding representations. In this paper, we first propose Neighbourhood-approximated Neural Theorem Provers (NaNTPs) consisting of two extensions to NTPs, namely a) a method for drastically reducing the previously prohibitive time and space complexity during inference and learning, and b) an attention mechanism for improving the rule learning process, deeming them usable on real-world datasets. Then, we propose a novel approach for jointly reasoning over KB facts and textual mentions, by jointly embedding them in a shared embedding space. The proposed method is able to extract rules and provide explanations—involving both textual patterns and KB relations—from large KBs and text corpora. We show that NaNTPs perform on par with NTPs at a fraction of a cost, and can achieve competitive link prediction results on challenging large-scale datasets, including WN18, WN18RR, and FB15k-237 (with and without textual mentions) while being able to provide explanations for each prediction and extract interpretable rules.

\*\*\*\*\*

Explaining Image Classifiers by Counterfactual Generation  
Chun-Hao Chang, Elliot Creager, Anna Goldenberg, David Duvenaud

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

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

y. 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.

\*\*\*\*\*

A Modern Take on the Bias-Variance Tradeoff in Neural Networks

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, Ioannis Mitliagkas

We revisit the bias-variance tradeoff for neural networks in light of modern empirical findings. The traditional bias-variance tradeoff in machine learning suggests that as model complexity grows, variance increases. Classical bounds in statistical learning theory point to the number of parameters in a model as a measure of model complexity, which means the tradeoff would indicate that variance increases with the size of neural networks. However, we empirically find that variance due to training set sampling is roughly constant (with both width and depth) in practice. Variance caused by the non-convexity of the loss landscape is different. We find that it decreases with width and increases with depth, in our setting. We provide theoretical analysis, in a simplified setting inspired by linear models, that is consistent with our empirical findings for width. We view bias-variance as a useful lens to study generalization through and encourage further theoretical explanation from this perspective.

\*\*\*\*\*

Towards a better understanding of Vector Quantized Autoencoders

Aurko Roy, Ashish Vaswani, Niki Parmar, Arvind Neelakantan

Deep neural networks with discrete latent variables offer the promise of better symbolic reasoning, and learning abstractions that are more useful to new tasks. There has been a surge in interest in discrete latent variable models, however, despite several recent improvements, the training of discrete latent variable models has remained challenging and their performance has mostly failed to match their continuous counterparts. Recent work on vector quantized autoencoders (VQ-VAE) has made substantial progress in this direction, with its perplexity almost matching that of a VAE on datasets such as CIFAR-10. In this work, we investigate an alternate training technique for VQ-VAE, inspired by its connection to the Expectation Maximization (EM) algorithm. Training the discrete autoencoder with EM and combining it with sequence level knowledge distillation allows us to develop a non-autoregressive machine translation model whose accuracy almost matches a strong greedy autoregressive baseline Transformer, while being 3.3 times faster at inference.

\*\*\*\*\*

Nested Dithered Quantization for Communication Reduction in Distributed Training

Afshin Abdi, Faramarz Fekri

In distributed training, the communication cost due to the transmission of gradients

or the parameters of the deep model is a major bottleneck in scaling up the number

of processing nodes. To address this issue, we propose dithered quantization for the transmission of the stochastic gradients and show that training with Dithered

Quantized Stochastic Gradients (DQSG) is similar to the training with unquantized

SGs perturbed by an independent bounded uniform noise, in contrast to the other quantization methods where the perturbation depends on the gradients and hence,

complicating the convergence analysis. We study the convergence of training algorithms using DQSG and the trade off between the number of quantization levels and the training time. Next, we observe that there is a correlation among the

SGs computed by workers that can be utilized to further reduce the communication overhead without any performance loss. Hence, we develop a simple yet effective quantization scheme, nested dithered quantized SG (NDQSG), that can reduce the communication significantly without requiring the workers communicating extra information to each other. We prove that although NDQSG requires significantly less bits, it can achieve the same quantization variance bound as DQSG. Our simulation results confirm the effectiveness of training using DQSG and NDQSG in reducing the communication bits or the convergence time compared to the existing methods without sacrificing the accuracy of the trained model.

\*\*\*\*\*

#### An Alarm System for Segmentation Algorithm Based on Shape Model

Fengze Liu, Yingda Xia, Dong Yang, Alan Yuille, Daguang Xu

It is usually hard for a learning system to predict correctly on the rare events, and there is no exception for segmentation algorithms. Therefore, we hope to build an alarm system to set off alarms when the segmentation result is possibly unsatisfactory. One plausible solution is to project the segmentation results into a low dimensional feature space, and then learn classifiers/regressors in the feature space to predict the qualities of segmentation results. In this paper, we form the feature space using shape feature which is a strong prior information shared among different data, so it is capable to predict the qualities of segmentation results given different segmentation algorithms on different datasets. The shape feature of a segmentation result is captured using the value of loss function when the segmentation result is tested using a Variational Auto-Encoder (VAE). The VAE is trained using only the ground truth masks, therefore the bad segmentation results with bad shapes become the rare events for VAE and will result in large loss value. By utilizing this fact, the VAE is able to detect all kinds of shapes that are out of the distribution of normal shapes in ground truth (GT). Finally, we learn the representation in the one-dimensional feature space to predict the qualities of segmentation results. We evaluate our alarm system on several recent segmentation algorithms for the medical segmentation task. The segmentation algorithms perform differently on different datasets, but our system consistently provides reliable prediction on the qualities of segmentation results.

\*\*\*\*\*

#### A Rate-Distortion Theory of Adversarial Examples

Angus Galloway, Anna Golubeva, Graham W. Taylor

The generalization ability of deep neural networks (DNNs) is intertwined with model complexity, robustness, and capacity. Through establishing an equivalence between a DNN and a noisy communication channel, we characterize generalization and fault tolerance for unbounded adversarial attacks in terms of information-theoretic quantities. Invoking rate-distortion theory, we suggest that excess capacity is a significant cause of vulnerability to adversarial examples.

\*\*\*\*\*

#### Security Analysis of Deep Neural Networks Operating in the Presence of Cache Side-Channel Attacks

Sanghyun Hong, Michael Davinroy, Yigitcan Kaya, Stuart Nevans Locke, Ian Rackow, Kevin Kulda, Dana Dachman-Soled, Tudor Dumitra

Recent work has introduced attacks that extract the architecture information of deep neural networks (DNN), as this knowledge enhances an adversary's capability to conduct attacks on black-box networks. This paper presents the first in-depth security analysis of DNN fingerprinting attacks that exploit cache side-channels. First, we define the threat model for these attacks: our adversary does not need the ability to query the victim model; instead, she runs a co-located process on the host machine victim's deep learning (DL) system is running and passively monitors the accesses of the target functions in the shared framework. S



second, we introduce DeepRecon, an attack that reconstructs the architecture of the victim network by using the internal information extracted via Flush+Reload, a cache side-channel technique. Once the attacker observes function invocations that map directly to architecture attributes of the victim network, the attacker can reconstruct the victim's entire network architecture. In our evaluation, we demonstrate that an attacker can accurately reconstruct two complex networks (VGG19 and ResNet50) having only observed one forward propagation. Based on the extracted architecture attributes, we also demonstrate that an attacker can build a meta-model that accurately fingerprints the architecture and family of the pre-trained model in a transfer learning setting. From this meta-model, we evaluate the importance of the observed attributes in the fingerprinting process. Third, we propose and evaluate new framework-level defense techniques that obfuscate our attacker's observations. Our empirical security analysis represents a step toward understanding the DNNs' vulnerability to cache side-channel attacks.

\*\*\*\*\*

#### Mode Normalization

Lucas Deecke, Iain Murray, Hakan Bilen

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

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

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 coefficient  $s$  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  $\times 1000$  speed-up compared to the previous work, which now allows for tractable data-driven metamer experiments.

\*\*\*\*\*

Neural MMO: A massively multiplayer game environment for intelligent agents  
Joseph Suarez, Yilun Du, Phillip Isola, Igor Mordatch

We present an artificial intelligence research platform inspired by the human game genre of MMORPGs (Massively Multiplayer Online Role-Playing Games, a.k.a. MMOs). We demonstrate how this platform can be used to study behavior and learning in large populations of neural agents. Unlike currently popular game environments, our platform supports persistent environments, with variable number of agents, and open-ended task descriptions. The emergence of complex life on Earth is often attributed to the arms race that ensued from a huge number of organisms all competing for finite resources. Our platform aims to simulate this setting in microcosm: we conduct a series of experiments to test how large-scale multiagent competition can incentivize the development of skillful behavior. We find that population size magnifies the complexity of the behaviors that emerge and results in agents that out-compete agents trained in smaller populations.

\*\*\*\*\*

NECST: Neural Joint Source-Channel Coding

Kristy Choi, Kedar Tatwawadi, Tsachy Weissman, Stefano Ermon

For reliable transmission across a noisy communication channel, classical results from information theory show that it is asymptotically optimal to separate out the source and channel coding processes. However, this decomposition can fall short in the finite bit-length regime, as it requires non-trivial tuning of hand-crafted codes and assumes infinite computational power for decoding. In this work, we propose Neural Error Correcting and Source Trimming (NECST) codes to jointly learn the encoding and decoding processes in an end-to-end fashion. By adding noise into the latent codes to simulate the channel during training, we learn to both compress and error-correct given a fixed bit-length and computational budget. We obtain codes that are not only competitive against several capacity-approaching channel codes, but also learn useful robust representations of the data for downstream tasks such as classification. Finally, we learn an extremely fast neural decoder, yielding almost an order of magnitude in speedup compared to standard decoding methods based on iterative belief propagation.

\*\*\*\*\*

BlackMarks: Black-box Multi-bit Watermarking for Deep Neural Networks

Huili Chen, Bitan Darvish Rouhani, Farinaz Koushanfar

Deep Neural Networks (DNNs) are increasingly deployed in cloud servers and autonomous agents due to their superior performance. The deployed DNN is either leveraged in a white-box setting (model internals are publicly known) or a black-box setting (only model outputs are known) depending on the application. A practical concern in the rush to adopt DNNs is protecting the models against Intellectual Property (IP) infringement. We propose BlackMarks, the first end-to-end multi-bit watermarking framework that is applicable in the black-box scenario. BlackMarks takes the pre-trained unmarked model and the owner's binary signature as inputs. The output is the corresponding marked model with specific keys that can be later used to trigger the embedded watermark. To do so, BlackMarks first designs a model-dependent encoding scheme that maps all possible classes in the task to bit '0' and bit '1'. Given the owner's watermark signature (a binary string), a set of key image and label pairs is designed using targeted adversarial attacks. The watermark (WM) is then encoded in the distribution of output activations of the DNN by fine-tuning the model with a WM-specific regularized loss. To extract the WM, BlackMarks queries the model with the WM key images and decodes the owner's signature from the corresponding predictions using the designed encoding

scheme. We perform a comprehensive evaluation of BlackMarks' performance on MNIST, CIFAR-10, ImageNet datasets and corroborate its effectiveness and robustness. BlackMarks preserves the functionality of the original DNN and incurs negligible WM embedding overhead as low as 2.054%.

\*\*\*\*\*

On Regularization and Robustness of Deep Neural Networks

Alberto Bietti\*, Grégoire Mialon\*, Julien Mairal

In this work, we study the connection between regularization and robustness of deep neural networks by viewing them as elements of a reproducing kernel Hilbert space (RKHS) of functions and by regularizing them using the RKHS norm. Even though this norm cannot be computed, we consider various approximations based on upper and lower bounds. These approximations lead to new strategies for regularization, but also to existing ones such as spectral norm penalties or constraints, gradient penalties, or adversarial training. Besides, the kernel framework allows us to obtain margin-based bounds on adversarial generalization. We show that our new algorithms lead to empirical benefits for learning on small datasets and learning adversarially robust models. We also discuss implications of our regularization framework for learning implicit generative models.

\*\*\*\*\*

Learning To Solve Circuit-SAT: An Unsupervised Differentiable Approach

Saeed Amizadeh, Sergiy Matushevych, Markus Weimer

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.

\*\*\*\*\*

Meta-Learning Neural Bloom Filters

Jack W Rae, Sergey Bartunov, Timothy P Lillicrap

There has been a recent trend in training neural networks to replace data structures that have been crafted by hand, with an aim for faster execution, better accuracy, or greater compression. In this setting, a neural data structure is instantiated by training a network over many epochs of its inputs until convergence. In many applications this expensive initialization is not practical, for example streaming algorithms --- where inputs are ephemeral and can only be inspected a small number of times. In this paper we explore the learning of approximate set membership over a stream of data in one-shot via meta-learning. We propose a novel memory architecture, the Neural Bloom Filter, which we show to be more compressive than Bloom Filters and several existing memory-augmented neural networks in scenarios of skewed data or structured sets.

\*\*\*\*\*

On Meaning-Preserving Adversarial Perturbations for Sequence-to-Sequence Models

Paul Michel, Graham Neubig, Xian Li, Juan Miguel Pino

Adversarial examples have been shown to be an effective way of assessing the robustness of neural sequence-to-sequence (seq2seq) models, by applying perturbations to the input of a model leading to large degradation in performance. However, these perturbations are only indicative of a weakness in the model if they do not change the semantics of the input in a way that would change the expected output. Using the example of machine translation (MT), we propose a new evaluation framework for adversarial attacks on seq2seq models taking meaning preservation into account and demonstrate that existing methods may not preserve meaning in general. Based on these findings, we propose new constraints for attacks on word-based MT systems and show, via human and automatic evaluation, that they produce more semantically similar adversarial inputs. Furthermore, we show that performing adversarial training with meaning-preserving attacks is beneficial to the model.

del in terms of adversarial robustness without hurting test performance.

\*\*\*\*\*

#### CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication

Nikita Kitaev, Jin-Hwa Kim, Xinlei Chen, Marcus Rohrbach, Yuandong Tian, Dhruv Batra, Devi Parikh

In this work, we propose a goal-driven collaborative task that contains language, vision, and action in a virtual environment as its core components. Specifically, we develop a Collaborative image-Drawing game between two agents, called CoDraw. Our game is grounded in a virtual world that contains movable clip art objects. The game involves two players: a Teller and a Drawer. The Teller sees an abstract scene containing multiple clip art pieces in a semantically meaningful configuration, while the Drawer tries to reconstruct the scene on an empty canvas using available clip art pieces. The two players communicate via two-way communication using natural language. We collect the CoDraw dataset of ~10K dialogs consisting of ~138K messages exchanged between human agents. We define protocols and metrics to evaluate the effectiveness of learned agents on this testbed, highlighting the need for a novel "crosstalk" condition which pairs agents trained independently on disjoint subsets of the training data for evaluation. We present models for our task, including simple but effective baselines and neural network approaches trained using a combination of imitation learning and goal-driven training. All models are benchmarked using both fully automated evaluation and by playing the game with live human agents.

\*\*\*\*\*

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

Hongzi Mao, Shaileshh Bojja Venkatakrisnan, Malte Schwarzkopf, Mohammad Alizadeh

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.

\*\*\*\*\*

#### Super-Resolution via Conditional Implicit Maximum Likelihood Estimation

Ke Li\*, Shichong Peng\*, Jitendra Malik

Single-image super-resolution (SISR) is a canonical problem with diverse applications. Leading methods like SRGAN produce images that contain various artifacts, such as high-frequency noise, hallucinated colours and shape distortions, which adversely affect the realism of the result. In this paper, we propose an alternative approach based on an extension of the method of Implicit Maximum Likelihood Estimation (IMLE). We demonstrate greater effectiveness at noise reduction and preservation of the original colours and shapes, yielding more realistic super-resolved images.

\*\*\*\*\*

#### Advocacy Learning

Ian Fox, Jenna Wiens

We introduce advocacy learning, a novel supervised training scheme for classification problems. This training scheme applies to a framework consisting of two connected networks: 1) the Advocates, composed of one subnetwork per class, which take the input and provide a convincing class-conditional argument in the form of an attention map, and 2) a Judge, which predicts the inputs class label based on these arguments. Each Advocate aims to convince the Judge that the input example belongs to their corresponding class. In contrast to a standard network, in

which all subnetworks are trained to jointly cooperate, we train the Advocates to competitively argue for their class, even when the input belongs to a different class. We also explore a variant, honest advocacy learning, where the Advocates are only trained on data corresponding to their class. Applied to several different classification tasks, we show that advocacy learning can lead to small improvements in classification accuracy over an identical supervised baseline. Through a series of follow-up experiments, we analyze when and how Advocates improve discriminative performance. Though it may seem counter-intuitive, a framework in which subnetworks are trained to competitively provide evidence in support of their class shows promise, performing as well as or better than standard approaches. This provides a foundation for further exploration into the effect of competition and class-conditional representations.

\*\*\*\*\*

#### Robustness May Be at Odds with Accuracy

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, Aleksander Madry

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

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.

\*\*\*\*\*

#### Riemannian TransE: Multi-relational Graph Embedding in Non-Euclidean Space

Atsushi Suzuki, Yosuke Enokida, Kenji Yamanishi

Multi-relational graph embedding which aims at achieving effective representations with reduced low-dimensional parameters, has been widely used in knowledge base completion. Although knowledge base data usually contains tree-like or cyclic structure, none of existing approaches can embed these data into a compatible space that in line with the structure. To overcome this problem, a novel framework, called Riemannian TransE, is proposed in this paper to embed the entities in a Riemannian manifold. Riemannian TransE models each relation as a move to a point and defines specific novel distance dissimilarity for each relation, so that

all the relations are naturally embedded in correspondence to the structure of data. Experiments on several knowledge base completion tasks have shown that, based on an appropriate choice of manifold, Riemannian TransE achieves good performance even with a significantly reduced parameters.

\*\*\*\*\*

#### Unlabeled Disentangling of GANs with Guided Siamese Networks

Gökhan Yildirim, Nikolay Jetchev, Urs Bergmann

Disentangling underlying generative factors of a data distribution is important for interpretability and generalizable representations. In this paper, we introduce two novel disentangling methods. Our first method, Unlabeled Disentangling GAN (UD-GAN, unsupervised), decomposes the latent noise by generating similar/dissimilar image pairs and it learns a distance metric on these pairs with siamese networks and a contrastive loss. This pairwise approach provides consistent representations for similar data points. Our second method (UD-GAN-G, weakly supervised) modifies the UD-GAN with user-defined guidance functions, which restrict the information that goes into the siamese networks. This constraint helps UD-GAN-G to focus on the desired semantic variations in the data. We show that both our methods outperform existing unsupervised approaches in quantitative metrics that measure semantic accuracy of the learned representations. In addition, we illustrate that simple guidance functions we use in UD-GAN-G allow us to directly capture the desired variations in the data.

\*\*\*\*\*

#### Link Prediction in Hypergraphs using Graph Convolutional Networks

Naganand Yadati, Vikram Nitin, Madhav Nimishakavi, Prateek Yadav, Anand Louis, Partha Talukdar

Link prediction in simple graphs is a fundamental problem in which new links between nodes are predicted based on the observed structure of the graph. However, in many real-world applications, there is a need to model relationships among nodes which go beyond pairwise associations. For example, in a chemical reaction, relationship among the reactants and products is inherently higher-order. Additionally, there is need to represent the direction from reactants to products. Hypergraphs provide a natural way to represent such complex higher-order relationships. Even though Graph Convolutional Networks (GCN) have recently emerged as a powerful deep learning-based approach for link prediction over simple graphs, their suitability for link prediction in hypergraphs is unexplored -- we fill this gap in this paper and propose Neural Hyperlink Predictor (NHP). NHP adapts GCNs for link prediction in hypergraphs. We propose two variants of NHP -- NHP-U and NHP-D -- for link prediction over undirected and directed hypergraphs, respectively. To the best of our knowledge, NHP-D is the first method for link prediction over directed hypergraphs. Through extensive experiments on multiple real-world datasets, we show NHP's effectiveness.

\*\*\*\*\*

#### BA-Net: Dense Bundle Adjustment Networks

Chengzhou Tang, Ping Tan

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

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.

\*\*\*\*\*

#### Heated-Up Softmax Embedding

Xu Zhang, Felix Xinnan Yu, Svebor Karaman, Wei Zhang, Shih-Fu Chang

Metric learning aims at learning a distance which is consistent with the semantic meaning of the samples. The problem is generally solved by learning an embedding, such that the samples of the same category are close (compact) while samples from different categories are far away (spread-out) in the embedding space. One popular way of generating such embeddings is to use the second-to-last layer of a deep neural network trained as a classifier with the softmax cross-entropy loss. In this paper, we show that training classifiers with different temperatures of the softmax function lead to different distributions of the embedding space. And finding a balance between the compactness, 'spread-out' and the generalization ability of the feature is critical in metric learning. Leveraging these insights, we propose a 'heating-up' strategy to train a classifier with increasing temperatures. Extensive experiments show that the proposed method achieves state-of-the-art embeddings on a variety of metric learning benchmarks.

\*\*\*\*\*

#### Learning Neural PDE Solvers with Convergence Guarantees

Jun-Ting Hsieh, Shengjia Zhao, Stephan Eismann, Lucia Mirabella, Stefano Ermon

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.

\*\*\*\*\*

#### Generative model based on minimizing exact empirical Wasserstein distance

Akihiro Iohara, Takahito Ogawa, Toshiyuki Tanaka

Generative Adversarial Networks (GANs) are a very powerful framework for generative modeling. However, they are often hard to train, and learning of GANs often becomes unstable. Wasserstein GAN (WGAN) is a promising framework to deal with the instability problem as it has a good convergence property. One drawback of the WGAN is that it evaluates the Wasserstein distance in the dual domain, which requires some approximation, so that it may fail to optimize the true Wasserstein distance. In this paper, we propose evaluating the exact empirical optimal transport cost efficiently in the primal domain and performing gradient descent with

respect to its derivative to train the generator network. Experiments on the MNIST dataset show that our method is significantly stable to converge, and achieves the lowest Wasserstein distance among the WGAN variants at the cost of some sharpness of generated images. Experiments on the 8-Gaussian toy dataset show that better gradients for the generator are obtained in our method. In addition, the proposed method enables more flexible generative modeling than WGAN.

\*\*\*\*\*

#### Reinforced Imitation Learning from Observations

Konrad Zolna, Negar Rostamzadeh, Yoshua Bengio, Sungjin Ahn, Pedro O. Pinheiro

Imitation learning is an effective alternative approach to learn a policy when the reward function is sparse. In this paper, we consider a challenging setting where an agent has access to a sparse reward function and state-only expert observations. We propose a method which gradually balances between the imitation learning cost and the reinforcement learning objective. Built upon an existing imitation learning method, our approach works with state-only observations. We show, through navigation scenarios, that (i) an agent is able to efficiently leverage sparse rewards to outperform standard state-only imitation learning, (ii) it can learn a policy even when learner's actions are different from the expert, and (iii) the performance of the agent is not bounded by that of the expert due to the optimized usage of sparse rewards.

\*\*\*\*\*

#### Meta-learning with differentiable closed-form solvers

Luca Bertinetto, Joao F. Henriques, Philip Torr, Andrea Vedaldi

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.

\*\*\*\*\*

#### Zero-shot Learning for Speech Recognition with Universal Phonetic Model

Xinjian Li, Siddharth Dalmia, David R. Mortensen, Florian Metze, Alan W Black

There are more than 7,000 languages in the world, but due to the lack of training sets, only a small number of them have speech recognition systems. Multilingual speech recognition provides a solution if at least some audio training data is available. Often, however, phoneme inventories differ between the training languages and the target language, making this approach infeasible. In this work, we address the problem of building an acoustic model for languages with zero audio resources. Our model is able to recognize unseen phonemes in the target language, if only a small text corpus is available. We adopt the idea of zero-shot learning, and decompose phonemes into corresponding phonetic attributes such as vowel and consonant. Instead of predicting phonemes directly, we first predict distributions over phonetic attributes, and then compute phoneme distributions with a customized acoustic model. We extensively evaluate our English-trained model on 20 unseen languages, and find that on average, it achieves 9.9% better phone error rate over a traditional CTC based acoustic model trained on English.



\*\*\*\*\*

## Towards Decomposed Linguistic Representation with Holographic Reduced Representation

Jiaming Luo, Yuan Cao, Yonghui Wu

The vast majority of neural models in Natural Language Processing adopt a form of structureless distributed representations. While these models are powerful at making predictions, the representational form is rather crude and does not provide insights into linguistic structures. In this paper we introduce novel language models with representations informed by the framework of Holographic Reduced Representation (HRR). This allows us to inject structures directly into our word-level and chunk-level representations. Our analyses show that by using HRR as a structured compositional representation, our models are able to discover crude linguistic roles, which roughly resembles a classic division between syntax and semantics.

\*\*\*\*\*

## SIMILE: Introducing Sequential Information towards More Effective Imitation Learning

Yutong Bai, Lingxi Xie

Reinforcement learning (RL) is a metaheuristic aiming at teaching an agent to interact with an environment and maximizing the reward in a complex task. RL algorithms often encounter the difficulty in defining a reward function in a sparse solution space. Imitation learning (IL) deals with this issue by providing a few expert demonstrations, and then either mimicking the expert's behavior (behavioral cloning, BC) or recovering the reward function by assuming the optimality of the expert (inverse reinforcement learning, IRL). Conventional IL approaches formulate the agent policy by mapping one single state to a distribution over actions, which did not consider sequential information. This strategy can be less accurate especially in IL, a weakly supervised learning environment, especially when the number of expert demonstrations is limited.

This paper presents an effective approach named Sequential IMItation LEarning (SIMILE). The core idea is to introduce sequential information, so that an agent can refer to both the current state and past state-action pairs to make a decision. We formulate our approach into a recurrent model, and instantiate it using LSTM so as to fuse both long-term and short-term information. SIMILE is a generalized IL framework which is easily applied to BC and IRL, two major types of IL algorithms. Experiments are performed on several robot controlling tasks in OpenAI Gym. SIMILE not only achieves performance gain over the baseline approaches, but also enjoys the benefit of faster convergence and better stability of testing performance. These advantages verify a higher learning efficiency of SIMILE, and implies its potential applications in real-world scenarios, i.e., when the agent-environment interaction is more difficult and/or expensive.

\*\*\*\*\*

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

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, Andrew McCallum

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

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>.

\*\*\*\*\*

One-Shot High-Fidelity Imitation: Training Large-Scale Deep Nets with RL  
Tom Le Paine, Sergio Gomez, Ziyu Wang, Scott Reed, Yusuf Aytar, Tobias Pfaff, Matt Hoffman, Gabriel Barth-Maron, Serkan Cabi, David Budden, Nando de Freitas

Humans are experts at high-fidelity imitation -- closely mimicking a demonstration, often in one attempt. Humans use this ability to quickly solve a task instance, and to bootstrap learning of new tasks. Achieving these abilities in autonomous agents is an open problem. In this paper, we introduce an off-policy RL algorithm (MetaMimic) to narrow this gap. MetaMimic can learn both (i) policies for high-fidelity one-shot imitation of diverse novel skills, and (ii) policies that enable the agent to solve tasks more efficiently than the demonstrators. MetaMimic relies on the principle of storing all experiences in a memory and replaying these to learn massive deep neural network policies by off-policy RL. This paper introduces, to the best of our knowledge, the largest existing neural networks for deep RL and shows that larger networks with normalization are needed to achieve one-shot high-fidelity imitation on a challenging manipulation task. The results also show that both types of policy can be learned from vision, in spite of the task rewards being sparse, and without access to demonstrator actions.

\*\*\*\*\*

The Expressive Power of Gated Recurrent Units as a Continuous Dynamical System  
Ian D. Jordan, Piotr Aleksander Sokol, Il Memming Park

Gated recurrent units (GRUs) were inspired by the common gated recurrent unit, long short-term memory (LSTM), as a means of capturing temporal structure with less complex memory unit architecture. Despite their incredible success in tasks such as natural and artificial language processing, speech, video, and polyphonic music, very little is understood about the specific dynamic features representable in a GRU network. As a result, it is difficult to know a priori how successful a GRU-RNN will perform on a given data set. In this paper, we develop a new theoretical framework to analyze one and two dimensional GRUs as a continuous dynamical system, and classify the dynamical features obtainable with such system. We found rich repertoire that includes stable limit cycles over time (nonlinear oscillations), multi-stable state transitions with various topologies, and homoclinic orbits. In addition, we show that any finite dimensional GRU cannot precisely replicate the dynamics of a ring attractor, or more generally, any continuous attractor, and is limited to finitely many isolated fixed points in theory. These findings were then experimentally verified in two dimensions by means of time series prediction.

\*\*\*\*\*

I Know the Feeling: Learning to Converse with Empathy  
Hannah Rashkin, Eric Michael Smith, Margaret Li, Y-Lan Boureau

Beyond understanding what is being discussed, human communication requires an awareness of what someone is feeling. One challenge for dialogue agents is recognizing feelings in the conversation partner and replying accordingly, a key communicative skill that is trivial for humans. Research in this area is made difficult by the paucity of suitable publicly available datasets both for emotion and dialogues. This work proposes a new task for empathetic dialogue generation and EmpatheticDialogues, a dataset of 25k conversations grounded in emotional situations to facilitate training and evaluating dialogue systems. Our experiments indic

ate that dialogue models that use our dataset are perceived to be more empathetic by human evaluators, while improving on other metrics as well (e.g. perceived relevance of responses, BLEU scores), compared to models merely trained on large-scale Internet conversation data. We also present empirical comparisons of several ways to improve the performance of a given model by leveraging existing models or datasets without requiring lengthy re-training of the full model.

\*\*\*\*\*

#### Graph Transformer

Yuan Li,Xiaodan Liang,Zhiting Hu,Yinbo Chen,Eric P. Xing

Graph neural networks (GNN) have gained increasing research interests as a mean to the challenging goal of robust and universal graph learning. Previous GNNs have assumed single pre-fixed graph structure and permitted only local context encoding. This paper proposes a novel Graph Transformer (GTR) architecture that captures long-range dependency with global attention, and enables dynamic graph structures. In particular, GTR propagates features within the same graph structure via an intra-graph message passing, and transforms dynamic semantics across multi-domain graph-structured data (e.g. images, sequences, knowledge graphs) for multi-modal learning via an inter-graph message passing. Furthermore, GTR enables effective incorporation of any prior graph structure by weighted averaging of the prior and learned edges, which can be crucially useful for scenarios where prior knowledge is desired. The proposed GTR achieves new state-of-the-arts across three benchmark tasks, including few-shot learning, medical abnormality and disease classification, and graph classification. Experiments show that GTR is superior in learning robust graph representations, transforming high-level semantics across domains, and bridging between prior graph structure with automatic structure learning.

\*\*\*\*\*

#### DHER: Hindsight Experience Replay for Dynamic Goals

Meng Fang,Cheng Zhou,Bei Shi,Boqing Gong,Jia Xu,Tong Zhang

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 policies 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

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

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  $v_x$  are a local maxima of  $v$ , according to  $h$  and satisfy  $c(v_x)=1$ . Therefore,  $c$  and  $h$  provide training signals to each other: a point  $v_x$  in the vicinity of  $v_x$  satisfies  $c(v_x)=-1$  or is deemed by  $h$  to be lower in value than  $v_x$ . 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.

\*\*\*\*\*

Multi-task Learning with Gradient Communication

Pengfei Liu, Xuanjing Huang

In this paper, we describe a general framework to systematically analyze current neural models for multi-task learning, in which we find that existing models expect to disentangle features into different spaces while features learned in practice are still entangled in shared space, leaving potential hazards for other training or unseen tasks. We propose to alleviate this problem by incorporating a new inductive bias into the process of multi-task learning, that different tasks can communicate with each other not only by passing hidden variables but gradients explicitly. Experimentally, we evaluate proposed methods on three groups of tasks and two types of settings (\textsc{in-task} and \textsc{out-of-task}). Quantitative and qualitative results show their effectiveness.

\*\*\*\*\*

MLPrune: Multi-Layer Pruning for Automated Neural Network Compression

Wenyuan Zeng, Raquel Urtasun

Model compression can significantly reduce the computation and memory footprint of large neural networks. To achieve a good trade-off between model size and accuracy, popular compression techniques usually rely on hand-crafted heuristics and

require manually setting the compression ratio of each layer. This process is typically costly and suboptimal. In this paper, we propose a Multi-Layer Pruning method (MLPrune), which is theoretically sound, and can automatically decide appropriate compression ratios for all layers. Towards this goal, we use an efficient approximation of the Hessian as our pruning criterion, based on a Kronecker-factored Approximate Curvature method. We demonstrate the effectiveness of our method on several datasets and architectures, outperforming previous state-of-the-art by a large margin. Our experiments show that we can compress AlexNet and VGG16 by 25x without loss in accuracy on ImageNet. Furthermore, our method has much fewer hyper-parameters and requires no expert knowledge.

\*\*\*\*\*

The Nonlinearity Coefficient - Predicting Generalization in Deep Neural Networks  
George Philipp,Jaime G. Carbonell

For a long time, designing neural architectures that exhibit high performance was considered a dark art that required expert hand-tuning. One of the few well-known guidelines for architecture design is the avoidance of exploding or vanishing gradients. However, even this guideline has remained relatively vague and circumstantial, because there exists no well-defined, gradient-based metric that can be computed {\it before} training begins and can robustly predict the performance of the network {\it after} training is complete.

We introduce what is, to the best of our knowledge, the first such metric: the nonlinearity coefficient (NLC). Via an extensive empirical study, we show that the NLC, computed in the network's randomly initialized state, is a powerful predictor of test error and that attaining a right-sized NLC is essential for attaining an optimal test error, at least in fully-connected feedforward networks. The NLC is also conceptually simple, cheap to compute, and is robust to a range of confounders and architectural design choices that comparable metrics are not necessarily robust to. Hence, we argue the NLC is an important tool for architecture search and design, as it can robustly predict poor training outcomes before training even begins.

\*\*\*\*\*

Neural Logic Machines

Honghua Dong,Jiayuan Mao,Tian Lin,Chong Wang,Lihong Li,Denny Zhou

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.

\*\*\*\*\*

Neural Regression Tree

Wenbo Zhao,Shahan Ali Memon,Bhiksha Raj,Rita Singh

Regression-via-Classification (RvC) is the process of converting a regression problem to a classification one. Current approaches for RvC use ad-hoc discretization strategies and are suboptimal. We propose a neural regression tree model for RvC. In this model, we employ a joint optimization framework where we learn optimal discretization thresholds while simultaneously optimizing the features for each node in the tree. We empirically show the validity of our model by testing it on two challenging regression tasks where we establish the state of the art.

\*\*\*\*\*

Locally Linear Unsupervised Feature Selection

Guillaume DOQUET,Michèle SEBAG

The paper, interested in unsupervised feature selection, aims to retain the features best accounting for the local patterns in the data. The proposed approach, called Locally Linear Unsupervised Feature Selection, relies on a dimensionality reduction method to characterize such patterns; each feature is thereafter assessed according to its compliance w.r.t. the local patterns, taking inspiration from Locally Linear Embedding (Roweis and Saul, 2000). The experimental validation of the approach on the scikit-feature benchmark suite demonstrates its effectiveness compared to the state of the art.

\*\*\*\*\*

How Training Data Affect the Accuracy and Robustness of Neural Networks for Image Classification

Suhua Lei,Huan Zhang,Ke Wang,Zhendong Su

Recent work has demonstrated the lack of robustness of well-trained deep neural networks (DNNs) to adversarial examples. For example, visually indistinguishable perturbations, when mixed with an original image, can easily lead deep learning models to misclassifications. In light of a recent study on the mutual influence between robustness and accuracy over 18 different ImageNet models, this paper investigates how training data affect the accuracy and robustness of deep neural

networks. We conduct extensive experiments on four different datasets, including CIFAR-10, MNIST, STL-10, and Tiny ImageNet, with several representative neural networks. Our results reveal previously unknown phenomena that exist between the size of training data and characteristics of the resulting models. In particular, besides confirming that the model accuracy improves as the amount of training data increases, we also observe that the model robustness improves initially, but there exists a turning point after which robustness starts to decrease. How and when such turning points occur vary for different neural networks and different datasets.

\*\*\*\*\*

Defensive Quantization: When Efficiency Meets Robustness

Ji Lin, Chuang Gan, Song Han

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.

\*\*\*\*\*

D-GAN: Divergent generative adversarial network for positive unlabeled learning and counter-examples generation

Florent CHIARONI. Mohamed-Cherif RAHAL. Nicolas HUEBER. Frédéric DUFAUX.

Positive Unlabeled (PU) learning consists in learning to distinguish samples of our class of interest, the positive class, from the counter-examples, the negative class, by using positive labeled and unlabeled samples during the training. Recent approaches exploit the GANs abilities to address the PU learning problem by generating relevant counter-examples. In this paper, we propose a new GAN-based PU learning approach named Divergent-GAN (D-GAN). The key idea is to incorporate a standard Positive Unlabeled learning risk inside the GAN discriminator loss function. In this way, the discriminator can ask the generator to converge towards the unlabeled samples distribution while diverging from the positive samples distribution. This enables the generator convergence towards the unlabeled counter-examples distribution without using prior knowledge, while keeping the standard adversarial GAN architecture. In addition, we discuss normalization techniques in the context of the proposed framework. Experimental results show that the proposed approach overcomes previous GAN-based PU learning methods issues, and it globally outperforms two-stage state of the art PU learning performances in terms of stability and prediction on both simple and complex image datasets.

\*\*\*\*\*

Zero-training Sentence Embedding via Orthogonal Basis

Ziyi Yang, Chenguang Zhu, Weizhu Chen

We propose a simple and robust training-free approach for building sentence representations. Inspired by the Gram-Schmidt Process in geometric theory, we build an orthogonal basis of the subspace spanned by a word and its surrounding context in a sentence. We model the semantic meaning of a word in a sentence based on two aspects. One is its relatedness to the word vector subspace already spanned by its contextual words. The other is its novel semantic meaning which shall be introduced as a new basis vector perpendicular to this existing subspace. Following this motivation, we develop an innovative method based on orthogonal basis to combine pre-trained word embeddings into sentence representation. This approach requires zero training and zero parameters, along with efficient inference performance. We evaluate our approach on 11 downstream NLP tasks. Experimental results show that our model outperforms all existing zero-training alternatives in all the tasks and it is competitive to other approaches relying on either large amounts of labelled data or prolonged training time.

\*\*\*\*\*

#### Model-Agnostic Meta-Learning for Multimodal Task Distributions

Risto Vuorio, Shao-Hua Sun, Hexiang Hu, Joseph J. Lim

Gradient-based meta-learners such as MAML (Finn et al., 2017) are able to learn a meta-prior from similar tasks to adapt to novel tasks from the same distribution with few gradient updates. One important limitation of such frameworks is that they seek a common initialization shared across the entire task distribution, substantially limiting the diversity of the task distributions that they are able to learn from. In this paper, we augment MAML with the capability to identify tasks sampled from a multimodal task distribution and adapt quickly through gradient updates. Specifically, we propose a multimodal MAML algorithm that is able to modulate its meta-learned prior according to the identified task, allowing faster adaptation. We evaluate the proposed model on a diverse set of problems including regression, few-shot image classification, and reinforcement learning. The results demonstrate the effectiveness of our model in modulating the meta-learned prior in response to the characteristics of tasks sampled from a multimodal distribution.

\*\*\*\*\*

#### DON'T JUDGE A BOOK BY ITS COVER - ON THE DYNAMICS OF RECURRENT NEURAL NETWORKS

Doron Haviv, Alexander Rivkind, Omri Barak

To be effective in sequential data processing, Recurrent Neural Networks (RNNs) are required to keep track of past events by creating memories. Consequently RNNs are harder to train than their feedforward counterparts, prompting the developments of both dedicated units such as LSTM and GRU and of a handful of training tricks. In this paper, we investigate the effect of different training protocols on the representation of memories in RNN. While reaching similar performance for different protocols, RNNs are shown to exhibit substantial differences in their ability to generalize for unforeseen tasks or conditions. We analyze the dynamics of the network's hidden state, and uncover the reasons for this difference. Each memory is found to be associated with a nearly steady state of the dynamics whose speed predicts performance on unforeseen tasks and which we refer to as a 'slow point'. By tracing the formation of the slow points we are able to understand the origin of differences between training protocols. Our results show that multiple solutions to the same task exist but may rely on different dynamical mechanisms, and that training protocols can bias the choice of such solutions in an interpretable way.

\*\*\*\*\*

#### AIM: Adversarial Inference by Matching Priors and Conditionals

Hanbo Li, Yaqing Wang, Changyou Chen, Jing Gao

Effective inference for a generative adversarial model remains an important and challenging problem. We propose a novel approach, Adversarial Inference by Matching priors and conditionals (AIM), which explicitly matches prior and conditional distributions in both data and code spaces, and puts a direct constraint on the dependency structure of the generative model. We derive an equivalent form of the prior and conditional matching objective that can be optimized efficiently without any parametric assumption on the data. We validate the effectiveness of A

IM on the MNIST, CIFAR-10, and CelebA datasets by conducting quantitative and qualitative evaluations. Results demonstrate that AIM significantly improves both reconstruction and generation as compared to other adversarial inference models.  
\*\*\*\*\*

Dimensionality Reduction for Representing the Knowledge of Probabilistic Models  
Marc T Law, Jake Snell, Amir-massoud Farahmand, Raquel Urtasun, Richard S Zemel

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

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

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.

\*\*\*\*\*

Learning Representations of Categorical Feature Combinations via Self-Attention  
Chen Xu, Chengzhen Fu, Peng Jiang, Wenwu Ou

Self-attention has been widely used to model the sequential data and achieved re



markable results in many applications. Although it can be used to model dependencies without regard to positions of sequences, self-attention is seldom applied to non-sequential data. In this work, we propose to learn representations of multi-field categorical data in prediction tasks via self-attention mechanism, where features are orderless but have intrinsic relations over different fields. In most current DNN based models, feature embeddings are simply concatenated for further processing by networks. Instead, by applying self-attention to transform the embeddings, we are able to relate features in different fields and automatically learn representations of their combinations, which are known as the factors of many prevailing linear models. To further improve the effect of feature combination mining, we modify the original self-attention structure by restricting the similarity weight to have at most  $k$  non-zero values, which additionally regularizes the model. We experimentally evaluate the effectiveness of our self-attention model on non-sequential data. Across two click through rate prediction benchmark datasets, i.e., Cretio and Avazu, our model with top- $k$  restricted self-attention achieves the state-of-the-art performance. Compared with the vanilla MLP, the gain by adding self-attention is significantly larger than that by modifying the network structures, which most current works focus on.

\*\*\*\*\*

#### ATTACK GRAPH CONVOLUTIONAL NETWORKS BY ADDING FAKE NODES

Xiaoyun Wang, Joe Eaton, Cho-Jui Hsieh, Felix Wu

Graph convolutional networks (GCNs) have been widely used for classifying graph nodes in the semi-supervised setting.

Previous works have shown that GCNs are vulnerable to the perturbation on adjacency and feature matrices of existing nodes. However, it is unrealistic to change the connections of existing nodes in many applications, such as existing users in social networks. In this paper, we investigate methods attacking GCNs by adding fake nodes. A greedy algorithm is proposed to generate adjacency and feature matrices of fake nodes, aiming to minimize the classification accuracy on the existing ones. In addition, we introduce a discriminator to classify fake nodes from real nodes, and propose a Greedy-GAN algorithm to simultaneously update the discriminator and the attacker, to make fake nodes indistinguishable to the real ones. Our non-targeted attack decreases the accuracy of GCN down to 0.10, and our targeted attack reaches a success rate of 0.99 for attacking the whole datasets, and 0.94 on average for attacking a single node.

\*\*\*\*\*

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

Johannes Gasteiger, Aleksandar Bojchevski, Stephan Günnemann

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

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, t

he policy should identify the specific nuances of an environment and exploit the m 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.

\*\*\*\*\*

Learning Diverse Generations using Determinantal Point Processes

Mohamed Elfeki,Camille Couprie,Mohamed Elhoseiny

Generative models have proven to be an outstanding tool for representing high-dimensional probability distributions and generating realistic looking images. A fundamental characteristic of generative models is their ability to produce multi-modal outputs. However, while training, they are often susceptible to mode collapse, which means that the model is limited in mapping the input noise to only a few modes of the true data distribution. In this paper, we draw inspiration from Determinantal Point Process (DPP) to devise a generative model that alleviates mode collapse while producing higher quality samples. DPP is an elegant probabilistic measure used to model negative correlations within a subset and hence quantify its diversity. We use DPP kernel to model the diversity in real data as well as in synthetic data. Then, we devise a generation penalty term that encourages the generator to synthesize data with a similar diversity to real data. In contrast to previous state-of-the-art generative models that tend to use additional trainable parameters or complex training paradigms, our method does not change the original training scheme. Embedded in an adversarial training and variational autoencoder, our Generative DPP approach shows a consistent resistance to mode-collapse on a wide-variety of synthetic data and natural image datasets including MNIST, CIFAR10, and CelebA, while outperforming state-of-the-art methods for data-efficiency, convergence-time, and generation quality. Our code will be made publicly available.

\*\*\*\*\*

Maximum a Posteriori on a Submanifold: a General Image Restoration Method with GAN

Fangzhou Luo,Xiaolin Wu

We propose a general method for various image restoration problems, such as denoising, deblurring, super-resolution and inpainting. The problem is formulated as a constrained optimization problem. Its objective is to maximize a posteriori probability of latent variables, and its constraint is that the image generated by these latent variables must be the same as the degraded image. We use a Generative Adversarial Network (GAN) as our density estimation model. Convincing results are obtained on MNIST dataset.

\*\*\*\*\*

Multi-objective training of Generative Adversarial Networks with multiple discriminators

Isabela Albuquerque,João Monteiro,Thang Doan,Breandan Considine,Tiago Falk,Ioannis Mitliagkas

Recent literature has demonstrated promising results on the training of Generative Adversarial Networks by employing a set of discriminators, as opposed to the traditional game involving one generator against a single adversary. Those methods perform single-objective optimization on some simple consolidation of the losses, e.g. an average. In this work, we revisit the multiple-discriminator approach by framing the simultaneous minimization of losses provided by different models as a multi-objective optimization problem. Specifically, we evaluate the performance of multiple gradient descent and the hypervolume maximization algorithm on a number of different datasets. Moreover, we argue that the previously proposed methods and hypervolume maximization can all be seen as variations of multiple

e gradient descent in which the update direction computation can be done efficiently. Our results indicate that hypervolume maximization presents a better compromise between sample quality and diversity, and computational cost than previous methods.

\*\*\*\*\*

#### Universal Successor Features for Transfer Reinforcement Learning

Chen Ma, Dylan R. Ashley, Junfeng Wen, Yoshua Bengio

Transfer in Reinforcement Learning (RL) refers to the idea of applying knowledge gained from previous tasks to solving related tasks. Learning a universal value function (Schaul et al., 2015), which generalizes over goals and states, has previously been shown to be useful for transfer. However, successor features are believed to be more suitable than values for transfer (Dayan, 1993; Barreto et al., 2017), even though they cannot directly generalize to new goals. In this paper, we propose (1) Universal Successor Features (USFs) to capture the underlying dynamics of the environment while allowing generalization to unseen goals and (2) a flexible end-to-end model of USFs that can be trained by interacting with the environment. We show that learning USFs is compatible with any RL algorithm that learns state values using a temporal difference method. Our experiments in a simple gridworld and with two MuJoCo environments show that USFs can greatly accelerate training when learning multiple tasks and can effectively transfer knowledge to new tasks.

\*\*\*\*\*

#### Transferring Knowledge across Learning Processes

Sebastian Flennerhag, Pablo G. Moreno, Neil D. Lawrence, Andreas Damianou

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 a 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.

\*\*\*\*\*

#### Latent Domain Transfer: Crossing modalities with Bridging Autoencoders

Yingtao Tian, Jesse Engel

Domain transfer is an exciting and challenging branch of machine learning because models must learn to smoothly transfer between domains, preserving local variations and capturing many aspects of variation without labels.

However, most successful applications to date require the two domains to be closely related (ex. image-to-image, video-video), utilizing similar or shared networks to transform domain specific properties like texture, coloring, and line shapes.

Here, we demonstrate that it is possible to transfer across modalities (ex. image-to-audio) by first abstracting the data with latent generative models and then learning transformations between latent spaces.

We find that a simple variational autoencoder is able to learn a shared latent space to bridge between two generative models in an unsupervised fashion, and even between different types of models (ex. variational autoencoder and a generative adversarial network).

We can further impose desired semantic alignment of attributes with a linear classifier in the shared latent space.

The proposed variation autoencoder enables preserving both locality and semantic alignment through the transfer process, as shown in the qualitative and quantitative evaluations.

Finally, the hierarchical structure decouples the cost of training the base gene

rative models and semantic alignments, enabling computationally efficient and data efficient retraining of personalized mapping functions.

\*\*\*\*\*

Time-Agnostic Prediction: Predicting Predictable Video Frames

Dinesh Jayaraman, Frederik Ebert, Alexei Efros, Sergey Levine

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

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 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

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

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

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

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.

\*\*\*\*\*

## Noisy Information Bottlenecks for Generalization

Julius Kunze, Louis Kirsch, Hippolyt Ritter, David Barber

We propose Noisy Information Bottlenecks (NIB) to limit mutual information between learned parameters and the data through noise. We show why this benefits generalization and allows mitigation of model overfitting both for supervised and unsupervised learning, even for arbitrarily complex architectures. We reinterpret methods including the Variational Autoencoder, beta-VAE, network weight uncertainty and a variant of dropout combined with weight decay as special cases of our approach, explaining and quantifying regularizing properties and vulnerabilities within information theory.

\*\*\*\*\*

## Open-Ended Content-Style Recombination Via Leakage Filtering

Karl Ridgeway, Michael C. Mozer

We consider visual domains in which a class label specifies the content of an image, and class-irrelevant properties that differentiate instances constitute the style. We present a domain-independent method that permits the open-ended recombination of style of one image with the content of another. Open ended simply means that the method generalizes to style and content not present in the training data. The method starts by constructing a content embedding using an existing deep metric-learning technique. This trained content encoder is incorporated into a variational autoencoder (VAE), paired with a to-be-trained style encoder. The VAE reconstruction loss alone is inadequate to ensure a decomposition of the latent representation into style and content. Our method thus includes an auxiliary loss, leakage filtering, which ensures that no style information remaining in the content representation is used for reconstruction and vice versa. We synthesize novel images by decoding the style representation obtained from one image with the content representation from another. Using this method for data-set augmentation, we obtain state-of-the-art performance on few-shot learning tasks.

\*\*\*\*\*

## Training Hard-Threshold Networks with Combinatorial Search in a Discrete Target Propagation Setting

Lukas Nabergall, Justin Toth, Leah Cousins

Learning deep neural networks with hard-threshold activation has recently become an important problem due to the proliferation of resource-constrained computing devices. In order to circumvent the inability to train with backpropagation in the presence of hard-threshold activations, \cite{friesen2017} introduced a discrete target propagation framework for training hard-threshold networks in a layer-by-layer fashion. Rather than using a gradient-based target heuristic, we explore the use of search methods for solving the target setting problem. Building on both traditional combinatorial optimization algorithms and gradient-based techniques, we develop a novel search algorithm Guided Random Local Search (GRLS). We demonstrate the effectiveness of our algorithm in training small networks on several datasets and evaluate our target-setting algorithm compared to simpler search methods and gradient-based techniques. Our results indicate that combinatorial optimization is a viable method for training hard-threshold networks that may have the potential to eventually surpass gradient-based methods in many settings.

\*\*\*\*\*

## Dimension-Free Bounds for Low-Precision Training

Zheng Li, Christopher De Sa

Low-precision training is a promising way of decreasing the time and energy cost of training machine learning models.

Previous work has analyzed low-precision training algorithms, such as low-precision stochastic gradient descent, and derived theoretical bounds on their convergence rates.

These bounds tend to depend on the dimension of the model  $d$  in that the number of bits needed to achieve a particular error bound increases as  $d$  increases. This is undesirable because a motivating application for low-precision training is large-scale models, such as deep learning, where  $d$  can be huge.

In this paper, we prove dimension-independent bounds for low-precision training

algorithms that use fixed-point arithmetic, which lets us better understand what affects the convergence of these algorithms as parameters scale.

Our methods also generalize naturally to let us prove new convergence bounds on low-precision training with other quantization schemes, such as low-precision floating-point computation and logarithmic quantization.

\*\*\*\*\*

VHEGAN: Variational Hetero-Encoder Randomized GAN for Zero-Shot Learning

Hao Zhang, Bo Chen, Long Tian, Zhengjue Wang, Mingyuan Zhou

To extract and relate visual and linguistic concepts from images and textual descriptions for text-based zero-shot learning (ZSL), we develop variational hetero-encoder (VHE) that decodes text via a deep probabilistic topic model, the variational posterior of whose local latent variables is encoded from an image via a Weibull distribution based inference network. To further improve VHE and add an image generator, we propose VHE randomized generative adversarial net (VHEGAN) that exploits the synergy between VHE and GAN through their shared latent space. After training with a hybrid stochastic-gradient MCMC/variational inference/stochastic gradient descent inference algorithm, VHEGAN can be used in a variety of settings, such as text generation/retrieval conditioning on an image, image generation/retrieval conditioning on a document/image, and generation of text-image pairs. The efficacy of VHEGAN is demonstrated quantitatively with experiments on both conventional and generalized ZSL tasks, and qualitatively on (conditional) image and/or text generation/retrieval.

\*\*\*\*\*

Theoretical Analysis of Auto Rate-Tuning by Batch Normalization

Sanjeev Arora, Zhiyuan Li, Kaifeng Lyu

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

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.

\*\*\*\*\*

Do Language Models Have Common Sense?

Trieu H. Trinh, Quoc V. Le

It has been argued that current machine learning models do not have commonsense, and therefore must be hard-coded with prior knowledge (Marcus, 2018). Here we s

how surprising evidence that language models can already learn to capture certain common sense knowledge. Our key observation is that a language model can compute the probability of any statement, and this probability can be used to evaluate the truthfulness of that statement. On the Winograd Schema Challenge (Levesque et al., 2011), language models are 11% higher in accuracy than previous state-of-the-art supervised methods. Language models can also be fine-tuned for the task of Mining Commonsense Knowledge on ConceptNet to achieve an F1 score of 0.912 and 0.824, outperforming previous best results (Jastrzebski et al., 2018). Further analysis demonstrates that language models can discover unique features of Winograd Schema contexts that decide the correct answers without explicit supervision.

\*\*\*\*\*

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

Wieland Brendel, Matthias Bethge

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 x 16 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.

\*\*\*\*\*

Predictive Local Smoothness for Stochastic Gradient Methods

Jun Li, Hongfu Liu, Bineng Zhong, Yue Wu, Yun Fu

Stochastic gradient methods are dominant in nonconvex optimization especially for deep models but have low asymptotical convergence due to the fixed smoothness. To address this problem, we propose a simple yet effective method for improving stochastic gradient methods named predictive local smoothness (PLS). First, we create a convergence condition to build a learning rate varied adaptively with local smoothness. Second, the local smoothness can be predicted by the latest gradients. Third, we use the adaptive learning rate to update the stochastic gradients for exploring linear convergence rates. By applying the PLS method, we implement new variants of three popular algorithms: PLS-stochastic gradient descent (PLS-SGD), PLS-accelerated SGD (PLS-AccSGD), and PLS-AMSGrad. Moreover, we provide much simpler proofs to ensure their linear convergence. Empirical results show that our variants have better performance gains than the popular algorithms, such as, faster convergence and alleviating explosion and vanish of gradients.

\*\*\*\*\*

Attentive Task-Agnostic Meta-Learning for Few-Shot Text Classification

Xiang Jiang, Mohammad Havaei, Gabriel Chartrand, Hassan Chouaib, Thomas Vincent, Andrew Jesson, Nicolas Chapados, Stan Matwin

Current deep learning based text classification methods are limited by their ability to achieve fast learning and generalization when the data is scarce. We address this problem by integrating a meta-learning procedure that uses the knowledge learned across many tasks as an inductive bias towards better natural language understanding. Inspired by the Model-Agnostic Meta-Learning framework (MAML), we introduce the Attentive Task-Agnostic Meta-Learning (ATAML) algorithm for text classification. The proposed ATAML is designed to encourage task-agnostic representation learning by way of task-agnostic parameterization and facilitate task-specific adaptation via attention mechanisms. We provide evidence to show that



the attention mechanism in ATAML has a synergistic effect on learning performance. Our experimental results reveal that, for few-shot text classification tasks, gradient-based meta-learning approaches outperform popular transfer learning methods. In comparisons with models trained from random initialization, pretrained models and meta trained MAML, our proposed ATAML method generalizes better on single-label and multi-label classification tasks in miniRCV1 and miniReuters-21578 datasets.

\*\*\*\*\*

#### Learning Exploration Policies for Navigation

Tao Chen, Saurabh Gupta, Abhinav Gupta

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

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.

\*\*\*\*\*

#### HR-TD: A Regularized TD Method to Avoid Over-Generalization

Ishan Durugkar, Bo Liu, Peter Stone

Temporal Difference learning with function approximation has been widely used recently and has led to several successful results. However, compared with the original tabular-based methods, one major drawback of temporal difference learning with neural networks and other function approximators is that they tend to over-generalize across temporally successive states, resulting in slow convergence and even instability. In this work, we propose a novel TD learning method, Hadamard product Regularized TD (HR-TD), that reduces over-generalization and thus leads to faster convergence. This approach can be easily applied to both linear and nonlinear function approximators.

HR-TD is evaluated on several linear and nonlinear benchmark domains, where we s

how improvement in learning behavior and performance.

\*\*\*\*\*

Learning Programmatically Structured Representations with Perceptor Gradients  
Svetlin Penkov, Subramanian Ramamoorthy

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.

\*\*\*\*\*

Mimicking actions is a good strategy for beginners: Fast Reinforcement Learning with Expert Action Sequences

Tharun Medini, Anshumali Shrivastava

Imitation Learning is the task of mimicking the behavior of an expert player in a Reinforcement Learning (RL) Environment to enhance the training of a fresh agent (called novice) beginning from scratch. Most of the Reinforcement Learning environments are stochastic in nature, i.e., the state sequences that an agent may encounter usually follow a Markov Decision Process (MDP). This makes the task of mimicking difficult as it is very unlikely that a new agent may encounter same or similar state sequences as an expert. Prior research in Imitation Learning proposes various ways to learn a mapping between the states encountered and the respective actions taken by the expert while mostly being agnostic to the order in which these were performed. Most of these methods need considerable number of states-action pairs to achieve good results. We propose a simple alternative to Imitation Learning by appending the novice's action space with the frequent short action sequences that the expert has taken. This simple modification, surprisingly improves the exploration and significantly outperforms alternative approaches like Dataset Aggregation. We experiment with several popular Atari games and show significant and consistent growth in the score that the new agents achieve using just a few expert action sequences.

\*\*\*\*\*

Learning Mixed-Curvature Representations in Product Spaces

Albert Gu, Frederic Sala, Beliz Gunel, Christopher Ré

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 i

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.

\*\*\*\*\*

#### Lyapunov-based Safe Policy Optimization

Yinlam Chow, Ofir Nachum, Mohammad Ghavamzadeh, Edgar Guzman-Duenez

In many reinforcement learning applications, it is crucial that the agent interacts with the environment only through safe policies, i.e., policies that do not take the agent to certain undesirable situations. These problems are often formulated as a constrained Markov decision process (CMDP) in which the agent's goal is to optimize its main objective while not violating a number of safety constraints. In this paper, we propose safe policy optimization algorithms that are based on the Lyapunov approach to CMDPs, an approach that has well-established theoretical guarantees in control engineering. We first show how to generate a set of state-dependent Lyapunov constraints from the original CMDP safety constraints. We then propose safe policy gradient algorithms that train a neural network policy using DDPG or PPO, while guaranteeing near-constraint satisfaction at every policy update by projecting either the policy parameter or the action onto the set of feasible solutions induced by the linearized Lyapunov constraints. Unlike the existing (safe) constrained PG algorithms, ours are more data efficient as they are able to utilize both on-policy and off-policy data. Furthermore, the action-projection version of our algorithms often leads to less conservative policy updates and allows for natural integration into an end-to-end PG training pipeline. We evaluate our algorithms and compare them with CPO and the Lagrangian method on several high-dimensional continuous state and action simulated robot locomotion tasks, in which the agent must satisfy certain safety constraints while minimizing its expected cumulative cost.

\*\*\*\*\*

#### Predictive Uncertainty through Quantization

Bastiaan S. Veeling, Rianne van den Berg, Max Welling

High-risk domains require reliable confidence estimates from predictive models. Deep latent variable models provide these, but suffer from the rigid variational distributions used for tractable inference, which err on the side of overconfidence.

We propose Stochastic Quantized Activation Distributions (SQUAD), which imposes a flexible yet tractable distribution over discretized latent variables. The proposed method is scalable, self-normalizing and sample efficient. We demonstrate that the model fully utilizes the flexible distribution, learns interesting non-linearities, and provides predictive uncertainty of competitive quality.

\*\*\*\*\*

#### Stable Recurrent Models

John Miller, Moritz Hardt

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.

\*\*\*\*\*

#### Inter-BMV: Interpolation with Block Motion Vectors for Fast Semantic Segmentation on Video

Samvit Jain, Joseph Gonzalez

Models optimized for accuracy on single images are often prohibitively slow to run on each frame in a video, especially on challenging dense prediction tasks, such as semantic segmentation. Recent work exploits the use of optical flow to warp image features forward from select keyframes, as a means to conserve computation on video. This approach, however, achieves only limited speedup, even when optimized, due to the accuracy degradation introduced by repeated forward warping, and the inference cost of optical flow estimation. To address these problems, we propose a new scheme that propagates features using the block motion vectors (BMV) present in compressed video (e.g. H.264 codecs), instead of optical flow, and bi-directionally warps and fuses features from enclosing keyframes to capture scene context on each video frame. Our technique, interpolation-BMV, enables us to accurately estimate the features of intermediate frames, while keeping inference costs low. We evaluate our system on the CamVid and Cityscapes datasets, comparing to both a strong single-frame baseline and related work. We find that we are able to substantially accelerate segmentation on video, achieving near real-time frame rates (20+ frames per second) on large images (e.g. 960 x 720 pixels), while maintaining competitive accuracy. This represents an improvement of almost 6x over the single-frame baseline and 2.5x over the fastest prior work.

\*\*\*\*\*

Success at any cost: value constrained model-free continuous control  
 Steven Bohez, Abbas Abdolmaleki, Michael Neunert, Jonas Buchli, Nicolas Heess, Raia Hadsell

Naively applying Reinforcement Learning algorithms to continuous control problems -- such as locomotion and robot control -- to maximize task reward often results in policies which rely on high-amplitude, high-frequency control signals, known colloquially as bang-bang control. While such policies can implement the optimal solution, particularly in simulated systems, they are often not desirable for real world systems since bang-bang control can lead to increased wear and tear and energy consumption and tends to excite undesired second-order dynamics. To counteract this issue, multi-objective optimization can be used to simultaneously optimize both the reward and some auxiliary cost that discourages undesired (e.g. high-amplitude) control. In principle, such an approach can yield the sought after, smooth, control policies. It can, however, be hard to find the correct trade-off between cost and return that results in the desired behavior. In this paper we propose a new constraint-based approach which defines a lower bound on the return while minimizing one or more costs (such as control effort). We employ Lagrangian relaxation to learn both (a) the parameters of a control policy that satisfies the desired constraints and (b) the Lagrangian multipliers for the optimization. Moreover, we demonstrate policy optimization which satisfies constraints either in expectation or in a per-step fashion, and we learn a single conditional policy that is able to dynamically change the trade-off between return and cost. We demonstrate the efficiency of our approach using a number of continuous control benchmark tasks as well as a realistic, energy-optimized quadruped locomotion task.

\*\*\*\*\*

Learning Heuristics for Automated Reasoning through Reinforcement Learning  
 Gil Lederman, Markus N. Rabe, Edward A. Lee, Sanjit A. Seshia  
 We demonstrate how to learn efficient heuristics for automated reasoning algorithms through deep reinforcement learning. We focus on backtracking search algorithms for quantified Boolean logics, which already can solve formulas of impressive size - up to 100s of thousands of variables. The main challenge is to find a representation of these formulas that lends itself to making predictions in a scalable way. For challenging problems, the heuristic learned through our approach

reduces execution time by a factor of 10 compared to the existing handwritten heuristics.

\*\*\*\*\*

#### Unsupervised Disentangling Structure and Appearance

Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, Chen Change Loy

It is challenging to disentangle an object into two orthogonal spaces of structure and appearance since each can influence the visual observation in a different and unpredictable way. It is rare for one to have access to a large number of data to help separate the influences. In this paper, we present a novel framework to learn this disentangled representation in a completely unsupervised manner. We address this problem in a two-branch Variational Autoencoder framework. For the structure branch, we project the latent factor into a soft structured point tensor and constrain it with losses derived from prior knowledge. This encourages the branch to distill geometry information. Another branch learns the complementary appearance information. The two branches form an effective framework that can disentangle object's structure-appearance representation without any human annotation. We evaluate our approach on four image datasets, on which we demonstrate the superior disentanglement and visual analogy quality both in synthesis and real-world data. We are able to generate photo-realistic images with 256\*256 resolution that are clearly disentangled in structure and appearance.

\*\*\*\*\*

#### Hallucinations in Neural Machine Translation

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, David Sussillo

Neural machine translation (NMT) systems have reached state of the art performance in translating text and are in wide deployment. Yet little is understood about how these systems function or break. Here we show that NMT systems are susceptible to producing highly pathological translations that are completely unrelated from the source material, which we term hallucinations. Such pathological translations are problematic because they are deeply disturbing of user trust and easy to find with a simple search. We describe a method to generate hallucinations and show that many common variations of the NMT architecture are susceptible to them. We study a variety of approaches to reduce the frequency of hallucinations, including data augmentation, dynamical systems and regularization techniques, showing that data augmentation significantly reduces hallucination frequency. Finally, we analyze networks that produce hallucinations and show that there are signatures in the attention matrix as well as in the hidden states of the decoder.

\*\*\*\*\*

#### Probabilistic Model-Based Dynamic Architecture Search

Nozomu Yoshinari, Kento Uchida, Shota Saito, Shinichi Shirakawa, Youhei Akimoto

The architecture search methods for convolutional neural networks (CNNs) have shown promising results. These methods require significant computational resources, as they repeat the neural network training many times to evaluate and search the architectures. Developing the computationally efficient architecture search method is an important research topic. In this paper, we assume that the structure parameters of CNNs are categorical variables, such as types and connectivities of layers, and they are regarded as the learnable parameters. Introducing the multivariate categorical distribution as the underlying distribution for the structure parameters, we formulate a differentiable loss for the training task, where the training of the weights and the optimization of the parameters of the distribution for the structure parameters are coupled. They are trained using the stochastic gradient descent, leading to the optimization of the structure parameters within a single training. We apply the proposed method to search the architecture for two computer vision tasks: image classification and inpainting. The experimental results show that the proposed architecture search method is fast and can achieve comparable performance to the existing methods.

\*\*\*\*\*

#### Neural Network Regression with Beta, Dirichlet, and Dirichlet-Multinomial Outputs

Peter Sadowski, Pierre Baldi

We propose a method for quantifying uncertainty in neural network regression models when the targets are real values on a  $d$ -dimensional simplex, such as probabilities. We show that each target can be modeled as a sample from a Dirichlet distribution, where the parameters of the Dirichlet are provided by the output of a neural network, and that the combined model can be trained using the gradient of the data likelihood. This approach provides interpretable predictions in the form of multidimensional distributions, rather than point estimates, from which one can obtain confidence intervals or quantify risk in decision making. Furthermore, we show that the same approach can be used to model targets in the form of empirical counts as samples from the Dirichlet-multinomial compound distribution. In experiments, we verify that our approach provides these benefits without harming the performance of the point estimate predictions on two diverse applications: (1) distilling deep convolutional networks trained on CIFAR-100, and (2) predicting the location of particle collisions in the XENON1T Dark Matter detector.

\*\*\*\*\*

Partially Mutual Exclusive Softmax for Positive and Unlabeled data

Ugo Tanielian, Flaviano Vasile, Mike Gartrell

In recent years, softmax together with its fast approximations has become the de-facto loss function for deep neural networks with multiclass predictions. However, softmax is used in many problems that do not fully fit the multiclass framework and where the softmax assumption of mutually exclusive outcomes can lead to biased results. This is often the case for applications such as language modeling, next event prediction and matrix factorization, where many of the potential outcomes are not mutually exclusive, but are more likely to be independent conditionally on the state. To this end, for the set of problems with positive and unlabeled data, we propose a relaxation of the original softmax formulation, where, given the observed state, each of the outcomes are conditionally independent but share a common set of negatives. Since we operate in a regime where explicit negatives are missing, we create an adversarially-trained model of negatives and derive a new negative sampling and weighting scheme which we denote as Cooperative Importance Sampling (CIS). We show empirically the advantages of our newly introduced negative sampling scheme by plugging it in the Word2Vec algorithm and benchmarking it extensively against other negative sampling schemes on both language modeling and matrix factorization tasks and show large lifts in performance.

\*\*\*\*\*

Automatic generation of object shapes with desired functionalities

Mihai Andries, Atabak Dehban, Jose Santos-Victor

3D objects (artefacts) are made to fulfill functions. Designing an object often starts with defining a list of functionalities that it should provide, also known as functional requirements. Today, the design of 3D object models is still a slow and largely artisanal activity, with few Computer-Aided Design (CAD) tools existing to aid the exploration of the design solution space. The purpose of the study is to explore the possibility of shape generation conditioned on desired functionalities. To accelerate the design process, we introduce an algorithm for generating object shapes with desired functionalities. We follow the principle form follows function, and assume that the form of a structure is correlated to its function. First, we use an artificial neural network to learn a function-to-form mapping by analysing a dataset of objects labeled with their functionalities. Then, we combine forms providing one or more desired functions, generating an object shape that is expected to provide all of them. Finally, we verify in simulation whether the generated object possesses the desired functionalities, by defining and executing functionality tests on it.

\*\*\*\*\*

Deep Anomaly Detection with Outlier Exposure

Dan Hendrycks, Mantas Mazeika, Thomas Dietterich

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.

\*\*\*\*\*

Variational recurrent models for representation learning

Qingming Tang, Mingda Chen, Weiran Wang, Karen Livescu

We study the problem of learning representations of sequence data. Recent work has built on variational autoencoders to develop variational recurrent models for generation. Our main goal is not generation but rather representation learning for downstream prediction tasks. Existing variational recurrent models typically use stochastic recurrent connections to model the dependence among neighboring latent variables, while generation assumes independence of generated data per time step given the latent sequence. In contrast, our models assume independence among all latent variables given non-stochastic hidden states, which speeds up inference, while assuming dependence of observations at each time step on all latent variables, which improves representation quality. In addition, we propose and study extensions for improving downstream performance, including hierarchical auxiliary latent variables and prior updating during training. Experiments show improved performance on several speech and language tasks with different levels of supervision, as well as in a multi-view learning setting.

\*\*\*\*\*

Trace-back along capsules and its application on semantic segmentation ■■

Tao Sun, Zhewei Wang, C. D. Smith, Jundong Liu

In this paper, we propose a capsule-based neural network model to solve the semantic segmentation problem. By taking advantage of the extractable part-whole dependencies available in capsule layers, we derive the probabilities of the class labels for individual capsules through a recursive, layer-by-layer procedure. We model this procedure as a traceback pipeline and take it as a central piece to build an end-to-end segmentation network. Under the proposed framework, image-level class labels and object boundaries are jointly sought in an explicit manner, which poses a significant advantage over the state-of-the-art fully convolutional network (FCN) solutions. Experiments conducted on modified MNIST and neuroimages demonstrate that our model considerably enhance the segmentation performance compared to the leading FCN variant.

\*\*\*\*\*

A Walk with SGD: How SGD Explores Regions of Deep Network Loss?

Chen Xing, Devansh Arpit, Christos Tsirigotis, Yoshua Bengio

The non-convex nature of the loss landscape of deep neural networks (DNN) lends them the intuition that over the course of training, stochastic optimization algorithms explore different regions of the loss surface by entering and escaping many local minima due to the noise induced by mini-batches. But is this really the case? This question couples the geometry of the DNN loss landscape with how stochastic optimization algorithms like SGD interact with it during training. Answering this question may help us qualitatively understand the dynamics of deep neural network optimization. We show evidence through qualitative and quantitative experiments that mini-batch SGD rarely crosses barriers during DNN optimization. As we show, the mini-batch induced noise helps SGD explore different regions of the loss surface using a seemingly different mechanism. To complement this finding, we also investigate the qualitative reason behind the slowing down of this exploration when using larger batch-sizes. We show this happens because gradients from larger batch-sizes align more with the top eigenvectors of the Hessian, which makes SGD oscillate in the proximity of the parameter initialization, thus

preventing exploration.

\*\*\*\*\*

## Perception-Aware Point-Based Value Iteration for Partially Observable Markov Decision Processes

Mahsa Ghasemi,Ufuk Topcu

Partially observable Markov decision processes (POMDPs) are a widely-used framework to model decision-making with uncertainty about the environment and under stochastic outcome. In conventional POMDP models, the observations that the agent receives originate from fixed known distribution. However, in a variety of real-world scenarios the agent has an active role in its perception by selecting which observations to receive. Due to combinatorial nature of such selection process, it is computationally intractable to integrate the perception decision with the planning decision. To prevent such expansion of the action space, we propose a greedy strategy for observation selection that aims to minimize the uncertainty in state.

We develop a novel point-based value iteration algorithm that incorporates the greedy strategy to achieve near-optimal uncertainty reduction for sampled belief points. This in turn enables the solver to efficiently approximate the reachable subspace of belief simplex by essentially separating computations related to perception from planning.

Lastly, we implement the proposed solver and demonstrate its performance and computational advantage in a range of robotic scenarios where the robot simultaneously performs active perception and planning.

\*\*\*\*\*

## Featurized Bidirectional GAN: Adversarial Defense via Adversarially Learned Semantic Inference

Ruying Bao,Sihang Liang,Qingcan Wang

Deep neural networks have been demonstrated to be vulnerable to adversarial attacks, where small perturbations intentionally added to the original inputs can fool the classifier. In this paper, we propose a defense method, Featurized Bidirectional Generative Adversarial Networks (FBGAN), to extract the semantic features of the input and filter the non-semantic perturbation. FBGAN is pre-trained on the clean dataset in an unsupervised manner, adversarially learning a bidirectional mapping between a high-dimensional data space and a low-dimensional semantic space; also mutual information is applied to disentangle the semantically meaningful features. After the bidirectional mapping, the adversarial data can be reconstructed to denoised data, which could be fed into any pre-trained classifier. We empirically show the quality of reconstruction images and the effectiveness of defense.

\*\*\*\*\*

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

Liqian Ma,Xu Jia,Stamatios Georgoulis,Tinne Tuytelaars,Luc Van Gool

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 insta



nces in the target domain, but also preserves the semantic consistency during the process.

\*\*\*\*\*

#### Clean-Label Backdoor Attacks

Alexander Turner, Dimitris Tsipras, Aleksander Madry

Deep neural networks have been recently demonstrated to be vulnerable to backdoor attacks. Specifically, by altering a small set of training examples, an adversary is able to install a backdoor that can be used during inference to fully control the model's behavior. While the attack is very powerful, it crucially relies on the adversary being able to introduce arbitrary, often clearly mislabeled, inputs to the training set and can thus be detected even by fairly rudimentary data filtering. In this paper, we introduce a new approach to executing backdoor attacks, utilizing adversarial examples and GAN-generated data. The key feature is that the resulting poisoned inputs appear to be consistent with their label and thus seem benign even upon human inspection.

\*\*\*\*\*

#### Adversarially Learned Mixture Model

Andrew Jesson, Cécale Low-Kam, Tanya Nair, Florian Soudan, Florent Chandelier, Nicolas Chapados

The Adversarially Learned Mixture Model (AMM) is a generative model for unsupervised or semi-supervised data clustering. The AMM is the first adversarially optimized method to model the conditional dependence between inferred continuous and categorical latent variables. Experiments on the MNIST and SVHN datasets show that the AMM allows for semantic separation of complex data when little or no labeled data is available. The AMM achieves unsupervised clustering error rates of 3.32% and 20.4% on the MNIST and SVHN datasets, respectively. A semi-supervised extension of the AMM achieves a classification error rate of 5.60% on the SVHN dataset.

\*\*\*\*\*

#### Inducing Cooperation via Learning to reshape rewards in semi-cooperative multi-agent reinforcement learning

David Earl Hostallero, Daewoo Kim, Kyunghwan Son, Yung Yi

We propose a deep reinforcement learning algorithm for semi-cooperative multi-agent tasks, where agents are equipped with their separate reward functions, yet with willingness to cooperate. Under these semi-cooperative scenarios, popular methods of centralized training with decentralized execution for inducing cooperation and removing the non-stationarity problem do not work well due to lack of a common shared reward as well as inscalability in centralized training. Our algorithm, called Peer-Evaluation based Dual DQN (PED-DQN), proposes to give peer evaluation signals to observed agents, which quantifies how they feel about a certain transition. This exchange of peer evaluation over time turns out to render agents to gradually reshape their reward functions so that their action choices from the myopic best-response tend to result in the good joint action with high cooperation. This evaluation-based method also allows flexible and scalable training by not assuming knowledge of the number of other agents and their observation and action spaces. We provide the performance evaluation of PED-DQN for the scenarios ranging from a simple two-person prisoner's dilemma to more complex semi-cooperative multi-agent tasks. In special cases where agents share a common reward function as in the centralized training methods, we show that inter-agent evaluation leads to better performance

\*\*\*\*\*

#### Pay Less Attention with Lightweight and Dynamic Convolutions

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, Michael Auli

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 cu

urrent 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

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 al. (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

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

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 a

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>.

\*\*\*\*\*

#### Dynamic Graph Representation Learning via Self-Attention Networks

Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, Hao Yang

Learning latent representations of nodes in graphs is an important and ubiquitous task with widespread applications such as link prediction, node classification, and graph visualization. Previous methods on graph representation learning mainly focus on static graphs, however, many real-world graphs are dynamic and evolve over time. In this paper, we present Dynamic Self-Attention Network (DySAT), a novel neural architecture that operates on dynamic graphs and learns node representations that capture both structural properties and temporal evolutionary patterns. Specifically, DySAT computes node representations by jointly employing self-attention layers along two dimensions: structural neighborhood and temporal dynamics. We conduct link prediction experiments on two classes of graphs: communication networks and bipartite rating networks. Our experimental results show that DySAT has a significant performance gain over several different state-of-the-art graph embedding baselines.

\*\*\*\*\*

#### Neural Networks with Structural Resistance to Adversarial Attacks

Luca de Alfaro

In adversarial attacks to machine-learning classifiers, small perturbations are added to input that is correctly classified. The perturbations yield adversarial examples, which are virtually indistinguishable from the unperturbed input, and yet are misclassified. In standard neural networks used for deep learning, attackers can craft adversarial examples from most input to cause a misclassification of their choice.

We introduce a new type of network units, called RBFI units, whose non-linear structure makes them inherently resistant to adversarial attacks. On permutation-invariant MNIST, in absence of adversarial attacks, networks using RBFI units match the performance of networks using sigmoid units, and are slightly below the accuracy of networks with ReLU units. When subjected to adversarial attacks based on projected gradient descent or fast gradient-sign methods, networks with RBFI units retain accuracies above 75%, while ReLU or Sigmoid see their accuracies reduced to below 1%.

Further, RBFI networks trained on regular input either exceed or closely match the accuracy of sigmoid and ReLU network trained with the help of adversarial examples.

The non-linear structure of RBFI units makes them difficult to train using standard gradient descent. We show that RBFI networks of RBFI units can be efficiently trained to high accuracies using pseudogradients, computed using functions especially crafted to facilitate learning instead of their true derivatives.

\*\*\*\*\*

#### Optimized Gated Deep Learning Architectures for Sensor Fusion

Myung Seok Shim, Peng Li

Sensor fusion is a key technology that integrates various sensory inputs to allow for robust decision making in many applications such as autonomous driving and robot control. Deep neural networks have been adopted for sensor fusion in a body of recent studies. Among these, the so-called netgated architecture was proposed, which has demonstrated improved performances over the conventional convolutional neural networks (CNN). In this paper, we address several limitations of the baseline negated architecture by proposing two further optimized architectures: a coarser-grained gated architecture employing (feature) group-level fusion weights and a two-stage gated architectures leveraging both the group-level and feature-level fusion weights. Using driving mode prediction and human activity

recognition datasets, we demonstrate the significant performance improvements brought by the proposed gated architectures and also their robustness in the presence of sensor noise and failures.

\*\*\*\*\*

#### NATTACK: A STRONG AND UNIVERSAL GAUSSIAN BLACK-BOX ADVERSARIAL ATTACK

Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, Boqing Gong

Recent works find that DNNs are vulnerable to adversarial examples, whose changes from the benign ones are imperceptible and yet lead DNNs to make wrong predictions. One can find various adversarial examples for the same input to a DNN using different attack methods. In other words, there is a population of adversarial examples, instead of only one, for any input to a DNN. By explicitly modeling this adversarial population with a Gaussian distribution, we propose a new black-box attack called NATTACK. The adversarial attack is hence formalized as an optimization problem, which searches the mean of the Gaussian under the guidance of increasing the target DNN's prediction error. NATTACK achieves 100% attack success rate on six out of eleven recently published defense methods (and greater than 90% for four), all using the same algorithm. Such results are on par with or better than powerful state-of-the-art white-box attacks. While the white-box attacks are often model-specific or defense-specific, the proposed black-box NATTACK is universally applicable to different defenses.

\*\*\*\*\*

#### ON THE EFFECTIVENESS OF TASK GRANULARITY FOR TRANSFER LEARNING

Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, Roland Memisevic

We describe a DNN for video classification and captioning, trained end-to-end, with shared features, to solve tasks at different levels of granularity, exploring the link between granularity in a source task and the quality of learned features for transfer learning. For solving the new task domain in transfer learning, we freeze the trained encoder and fine-tune an MLP on the target domain. We train on the Something-Something dataset with over 220,000 videos, and multiple levels of target granularity, including 50 action groups, 174 fine-grained action categories and captions. Classification and captioning with Something-Something are challenging because of the subtle differences between actions, applied to thousands of different object classes, and the diversity of captions penned by crowd actors. Our model performs better than existing classification baselines for Something-Something, with impressive fine-grained results. And it yields a strong baseline on the new Something-Something captioning task. Experiments reveal that training with more fine-grained tasks tends to produce better features for transfer learning.

\*\*\*\*\*

#### A RECURRENT NEURAL CASCADE-BASED MODEL FOR CONTINUOUS-TIME DIFFUSION PROCESS

Sylvain Lamprier

Many works have been proposed in the literature to capture the dynamics of diffusion in networks. While some of them define graphical markovian models to extract temporal relationships between node infections in networks, others consider diffusion episodes as sequences of infections via recurrent neural models. In this paper we propose a model at the crossroads of these two extremes, which embeds the history of diffusion in infected nodes as hidden continuous states. Depending on the trajectory followed by the content before reaching a given node, the distribution of influence probabilities may vary. However, content trajectories are usually hidden in the data, which induces challenging learning problems. We propose a topological recurrent neural model which exhibits good experimental per-

performances for diffusion modelling and prediction.

\*\*\*\*\*

Random mesh projectors for inverse problems

Konik Kothari\*, Sidharth Gupta\*, Maarten v. de Hoop, Ivan Dokmanic

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.

\*\*\*\*\*

Shrinkage-based Bias-Variance Trade-off for Deep Reinforcement Learning

Yihao Feng, Hao Liu, Jian Peng, Qiang Liu

Deep reinforcement learning has achieved remarkable successes in solving various challenging artificial intelligence tasks. A variety of different algorithms have been introduced and improved towards human-level performance. Although technical advances have been developed for each individual algorithms, there has been strong evidence showing that further substantial improvements can be achieved by properly combining multiple approaches with difference biases and variances. In this work, we propose to use the James-Stein (JS) shrinkage estimator to combine on-policy policy gradient estimators which have low bias but high variance, with low-variance high-bias gradient estimates such as those constructed based on model-based methods or temporally smoothed averaging of historical gradients. Empirical results show that our simple shrinkage approach is very effective in practice and substantially improve the sample efficiency of the state-of-the-art on-policy methods on various continuous control tasks.

\*\*\*\*\*

A Statistical Approach to Assessing Neural Network Robustness

Stefan Webb, Tom Rainforth, Yee Whye Teh, M. Pawan Kumar

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

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 gene

rative 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

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.

\*\*\*\*\*

## Understanding & Generalizing AlphaGo Zero

Ravichandra Addanki, Mohammad Alizadeh, Shaileshh Bojja Venkatakrishnan, Devavrat Shah, Qiaomin Xie, Zhi Xu

AlphaGo Zero (AGZ) introduced a new {\em tabula rasa} reinforcement learning algorithm that has achieved superhuman performance in the games of Go, Chess, and Shogi with no prior knowledge other than the rules of the game. This success naturally begs the question whether it is possible to develop similar high-performance reinforcement learning algorithms for generic sequential decision-making problems (beyond two-player games), using only the constraints of the environment as the ``rules.'' To address this challenge, we start by taking steps towards developing a formal understanding of AGZ. AGZ includes two key innovations: (1) it learns a policy (represented as a neural network) using {\em supervised learning} with cross-entropy loss from samples generated via Monte-Carlo Tree Search (MCTS); (2) it uses {\em self-play} to learn without training data.

We argue that the self-play in AGZ corresponds to learning a Nash equilibrium for the two-player game; and the supervised learning with MCTS is attempting to learn the policy corresponding to the Nash equilibrium, by establishing a novel bound on the difference between the expected return achieved by two policies in terms of the expected KL divergence (cross-entropy) of their induced distributions. To extend AGZ to generic sequential decision-making problems, we introduce a {\em robust MDP} framework, in which the agent and nature effectively play a zero-sum game: the agent aims to take actions to maximize reward while nature seeks state transitions, subject to the constraints of that environment, that minimize

the agent's reward. For a challenging network scheduling domain, we find that A GZ within the robust MDP framework provides near-optimal performance, matching one of the best known scheduling policies that has taken the networking community three decades of intensive research to develop.

\*\*\*\*\*

Quality Evaluation of GANs Using Cross Local Intrinsic Dimensionality  
Sukarna Barua,Xingjun Ma,Sarah Monazam Erfani,Michael Houle,James Bailey  
Generative Adversarial Networks (GANs) are an elegant mechanism for data generation. However, a key challenge when using GANs is how to best measure their ability to generate realistic data. In this paper, we demonstrate that an intrinsic dimensional characterization of the data space learned by a GAN model leads to an effective evaluation metric for GAN quality. In particular, we propose a new evaluation measure, CrossLID, that assesses the local intrinsic dimensionality (LID) of input data with respect to neighborhoods within GAN-generated samples. In experiments on 3 benchmark image datasets, we compare our proposed measure to several state-of-the-art evaluation metrics. Our experiments show that CrossLID is strongly correlated with sample quality, is sensitive to mode collapse, is robust to small-scale noise and image transformations, and can be applied in a model-free manner. Furthermore, we show how CrossLID can be used within the GAN training process to improve generation quality.

\*\*\*\*\*

Guided Evolutionary Strategies: Escaping the curse of dimensionality in random search

Niru Maheswaranathan,Luke Metz,George Tucker,Dami Choi,Jascha Sohl-Dickstein  
Many applications in machine learning require optimizing a function whose true gradient is unknown, but where surrogate gradient information (directions that may be correlated with, but not necessarily identical to, the true gradient) is available instead. This arises when an approximate gradient is easier to compute than the full gradient (e.g. in meta-learning or unrolled optimization), or when a true gradient is intractable and is replaced with a surrogate (e.g. in certain reinforcement learning applications or training networks with discrete variables). We propose Guided Evolutionary Strategies, a method for optimally using surrogate gradient directions along with random search. We define a search distribution for evolutionary strategies that is elongated along a subspace spanned by the surrogate gradients. This allows us to estimate a descent direction which can then be passed to a first-order optimizer. We analytically and numerically characterize the tradeoffs that result from tuning how strongly the search distribution is stretched along the guiding subspace, and use this to derive a setting of the hyperparameters that works well across problems. Finally, we apply our method to example problems including truncated unrolled optimization and training neural networks with discrete variables, demonstrating improvement over both standard evolutionary strategies and first-order methods (that directly follow the surrogate gradient). We provide a demo of Guided ES at: [redacted URL](#)

\*\*\*\*\*

Meta-Learning for Contextual Bandit Exploration

Amr Sharaf,Hal Daumé III

We describe MÊLÉE, a meta-learning algorithm for learning a good exploration policy in the interactive contextual bandit setting. Here, an algorithm must take actions based on contexts, and learn based only on a reward signal from the action taken, thereby generating an exploration/exploitation trade-off. MÊLÉE addresses this trade-off by learning a good exploration strategy based on offline synthetic tasks, on which it can simulate the contextual bandit setting. Based on these simulations, MÊLÉE uses an imitation learning strategy to learn a good exploration policy that can then be applied to true contextual bandit tasks at test time. We compare MÊLÉE to seven strong baseline contextual bandit algorithms on a set of three hundred real-world datasets, on which it outperforms alternatives in most settings, especially when differences in rewards are large. Finally, we demonstrate the importance of having a rich feature representation for learning how to explore.

\*\*\*\*\*

Reconciling Feature-Reuse and Overfitting in DenseNet with Specialized Dropout  
Kun Wan, Boyuan Feng, Lingwei Xie, Yufei Ding

Recently convolutional neural networks (CNNs) achieve great accuracy in visual recognition tasks. DenseNet becomes one of the most popular CNN models due to its effectiveness in feature-reuse. However, like other CNN models, DenseNets also face overfitting problem if not severer. Existing dropout method can be applied but not as effective due to the introduced nonlinear connections. In particular, the property of feature-reuse in DenseNet will be impeded, and the dropout effect will be weakened by the spatial correlation inside feature maps. To address these problems, we craft the design of a specialized dropout method from three aspects, dropout location, dropout granularity, and dropout probability. The insights attained here could potentially be applied as a general approach for boosting the accuracy of other CNN models with similar nonlinear connections. Experimental results show that DenseNets with our specialized dropout method yield better accuracy compared to vanilla DenseNet and state-of-the-art CNN models, and such accuracy boost increases with the model depth.

\*\*\*\*\*

Understanding the Effectiveness of Lipschitz-Continuity in Generative Adversarial Nets

Zhiming Zhou, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang, Zhihua Zhang, Yong Yu

In this paper, we investigate the underlying factor that leads to the failure and success in training of GANs. Specifically, we study the property of the optimal discriminative function  $f^*(x)$  and show that  $f^*(x)$  in most GANs can only reflect the local densities at  $x$ , which means the value of  $f^*(x)$  for points in the fake distribution ( $P_g$ ) does not contain any information useful about the location of other points in the real distribution ( $P_r$ ). Given that the supports of the real and fake distributions are usually disjoint, we argue that such a  $f^*(x)$  and its gradient tell nothing about "how to pull  $P_g$  to  $P_r$ ", which turns out to be the fundamental cause of failure in training of GANs. We further demonstrate that a well-defined distance metric (including the dual form of Wasserstein distance with a compacted constraint) does not necessarily ensure the convergence of GANs. Finally, we propose Lipschitz-continuity condition as a general solution and show that in a large family of GAN objectives, Lipschitz condition is capable of connecting  $P_g$  and  $P_r$  through  $f^*(x)$  such that the gradient  $\nabla_x f^*(x)$  at each sample  $x \sim P_g$  points towards some real sample  $y \sim P_r$ .

\*\*\*\*\*

On the Geometry of Adversarial Examples

Marc Khoury, Dylan Hadfield-Menell

Adversarial examples are a pervasive phenomenon of machine learning models where seemingly imperceptible perturbations to the input lead to misclassifications for otherwise statistically accurate models. We propose a geometric framework, drawing on tools from the manifold reconstruction literature, to analyze the high-dimensional geometry of adversarial examples. In particular, we highlight the importance of codimension: for low-dimensional data manifolds embedded in high-dimensional space there are many directions off the manifold in which to construct adversarial examples. Adversarial examples are a natural consequence of learning a decision boundary that classifies the low-dimensional data manifold well, but classifies points near the manifold incorrectly. Using our geometric framework we prove (1) a tradeoff between robustness under different norms, (2) that adversarial training in balls around the data is sample inefficient, and (3) sufficient sampling conditions under which nearest neighbor classifiers and ball-based adversarial training are robust.

\*\*\*\*\*

Transfer Learning for Sequences via Learning to Collocate

Wanyun Cui, Guangyu Zheng, Zhiqiang Shen, Sihang Jiang, Wei Wang

Transfer learning aims to solve the data sparsity for a specific domain by apply



ing 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.

\*\*\*\*\*

Modulated Variational Auto-Encoders for Many-to-Many Musical Timbre Transfer

Adrien Bitton, Philippe Esling, Axel Chemla-Romeu-Santos

Generative models have been successfully applied to image style transfer and domain translation. However, there is still a wide gap in the quality of results when learning such tasks on musical audio. Furthermore, most translation models only enable one-to-one or one-to-many transfer by relying on separate encoders or decoders and complex, computationally-heavy models. In this paper, we introduce the Modulated Variational auto-Encoders (MoVE) to perform musical timbre transfer. First, we define timbre transfer as applying parts of the auditory properties of a musical instrument onto another. We show that we can achieve and improve this task by conditioning existing domain translation techniques with Feature-wise Linear Modulation (FiLM). Then, by replacing the usual adversarial translation criterion by a Maximum Mean Discrepancy (MMD) objective, we alleviate the need for an auxiliary pair of discriminative networks. This allows a faster and more stable training, along with a controllable latent space encoder. By further conditioning our system on several different instruments, we can generalize to many-to-many transfer within a single variational architecture able to perform multi-domain transfers. Our models map inputs to 3-dimensional representations, successfully translating timbre from one instrument to another and supporting sound synthesis on a reduced set of control parameters. We evaluate our method in reconstruction and generation tasks while analyzing the auditory descriptor distributions across transferred domains. We show that this architecture incorporates generative controls in multi-domain transfer, yet remaining rather light, fast to train and effective on small datasets.

\*\*\*\*\*

SSoC: Learning Spontaneous and Self-Organizing Communication for Multi-Agent Collaboration

Xiangyu Kong, Jing Li, Bo Xin, Yizhou Wang

Multi-agent collaboration is required by numerous real-world problems. Although distributed setting is usually adopted by practical systems, local range communication and information aggregation still matter in fulfilling complex tasks. For multi-agent reinforcement learning, many previous studies have been dedicated to design an effective communication architecture. However, existing models usually suffer from an ossified communication structure, e.g., most of them predefine a particular communication mode by specifying a fixed time frequency and spatial scope for agents to communicate regardless of necessity. Such design is incapable of dealing with multi-agent scenarios that are capricious and complicated, especially when only partial information is available. Motivated by this, we argue that the solution is to build a spontaneous and self-organizing communication (SSoC) learning scheme. By treating the communication behaviour as an explicit action, SSoC learns to organize communication in an effective and efficient way.

Particularly, it enables each agent to spontaneously decide when and who to send messages based on its observed states. In this way, a dynamic inter-agent communication channel is established in an online and self-organizing manner. The agents also learn how to adaptively aggregate the received messages and its own hidden states to execute actions. Various experiments have been conducted to demonstrate that SSoC really learns intelligent message passing among agents located far apart. With such agile communications, we observe that effective collaboration tactics emerge which have not been mastered by the compared baselines.

\*\*\*\*\*

How to train your MAML

Antreas Antoniou, Harrison Edwards, Amos Storkey

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++.

\*\*\*\*\*

Meta Learning with Fast/Slow Learners

zhuoyuan@fb.com

Meta-learning has recently achieved success in many optimization problems. In general, a meta learner  $g(\cdot)$  could be learned for a base model  $f(\cdot)$  on a variety of tasks, such that it can be more efficient on a new task. In this paper, we make some key modifications to enhance the performance of meta-learning models. (1) we leverage different meta-strategies for different modules to optimize them separately: we use conservative "slow learners" on low-level basic feature representation layers and "fast learners" on high-level task-specific layers; (2) Furthermore, we provide theoretical analysis on why the proposed approach works, based on a case study on a two-layer MLP. We evaluate our model on synthetic MLP regression, as well as low-shot learning tasks on Omniglot and ImageNet benchmarks. We demonstrate that our approach is able to achieve state-of-the-art performance.

\*\*\*\*\*

A Study of Robustness of Neural Nets Using Approximate Feature Collisions

Ke Li\*, Tianhao Zhang\*, Jitendra Malik

In recent years, various studies have focused on the robustness of neural nets. While it is known that neural nets are not robust to examples with adversarially chosen perturbations as a result of linear operations on the input data, we show in this paper there could be a convex polytope within which all examples are misclassified by neural nets due to the properties of ReLU activation functions. We propose a way to find such polytopes empirically and demonstrate that such polytopes exist in practice. Furthermore, we show that such polytopes exist even after constraining the examples to be a composition of image patches, resulting in perceptibly different examples at different locations in the polytope that are all misclassified.

\*\*\*\*\*

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

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 bottle

neck. 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.

\*\*\*\*\*

#### Information Regularized Neural Networks

Tianchen Zhao, Dejiao Zhang, Zeyu Sun, Honglak Lee

We formulate an information-based optimization problem for supervised classification. For invertible neural networks, the control of these information terms is passed down to the latent features and parameter matrix in the last fully connected layer, given that mutual information is invariant under invertible map. We propose an objective function and prove that it solves the optimization problem.

Our framework allows us to learn latent features in a more interpretable form while improving the classification performance. We perform extensive quantitative and qualitative experiments in comparison with the existing state-of-the-art classification models.

\*\*\*\*\*

#### Prob2Vec: Mathematical Semantic Embedding for Problem Retrieval in Adaptive Tutoring

Du Su, Ali Yekkehkhany, Yi Lu, Wenmiao Lu

We propose a new application of embedding techniques to problem retrieval in adaptive tutoring. The objective is to retrieve problems similar in mathematical concepts. There are two challenges: First, like sentences, problems helpful to tutoring are never exactly the same in terms of the underlying concepts. Instead, good problems mix concepts in innovative ways, while still displaying continuity in their relationships. Second, it is difficult for humans to determine a similarity score consistent across a large enough training set. We propose a hierarchical problem embedding algorithm, called Prob2Vec, that consists of an abstraction and an embedding step. Prob2Vec achieves 96.88% accuracy on a problem similarity test, in contrast to 75% from directly applying state-of-the-art sentence embedding methods. It is surprising that Prob2Vec is able to distinguish very fine-grained differences among problems, an ability humans need time and effort to acquire. In addition, the sub-problem of concept labeling with imbalanced training data set is interesting in its own right. It is a multi-label problem suffering from dimensionality explosion, which we propose ways to ameliorate. We propose the novel negative pre-training algorithm that dramatically reduces false negative and positive ratios for classification, using an imbalanced training dataset.

\*\*\*\*\*

#### Learning what and where to attend

Drew Linsley, Dan Shiebler, Sven Eberhardt, Thomas Serre

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

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.

\*\*\*\*\*

Conditional Inference in Pre-trained Variational Autoencoders via Cross-coding

Ga Wu, Justin Domke, Scott Sanner

Variational Autoencoders (VAEs) are a popular generative model, but one in which conditional inference can be challenging. If the decomposition into query and evidence variables is fixed, conditional VAEs provide an attractive solution. To support arbitrary queries, one is generally reduced to Markov Chain Monte Carlo sampling methods that can suffer from long mixing times. In this paper, we propose an idea we term cross-coding to approximate the distribution over the latent variables after conditioning on an evidence assignment to some subset of the variables. This allows generating query samples without retraining the full VAE.

We experimentally evaluate three variations of cross-coding showing that (i) can be quickly optimized for different decompositions of evidence and query and (ii) they quantitatively and qualitatively outperform Hamiltonian Monte Carlo.

\*\*\*\*\*

Safe Policy Learning from Observations

Elad Sarafian, Aviv Tamar, Sarit Kraus

In this paper, we consider the problem of learning a policy by observing numerous non-expert agents. Our goal is to extract a policy that, with high-confidence, acts better than the agents' average performance. Such a setting is important for real-world problems where expert data is scarce but non-expert data can easily be obtained, e.g. by crowdsourcing. Our approach is to pose this problem as safe policy improvement in reinforcement learning. First, we evaluate an average behavior policy and approximate its value function. Then, we develop a stochastic policy improvement algorithm that safely improves the average behavior. The primary advantages of our approach, termed Rerouted Behavior Improvement (RBI), over other safe learning methods are its stability in the presence of value estimation errors and the elimination of a policy search process. We demonstrate these advantages in the Taxi grid-world domain and in four games from the Atari learning environment.

\*\*\*\*\*

## Open Vocabulary Learning on Source Code with a Graph-Structured Cache

Milan Cvitkovic,Badal Singh,Anima Anandkumar

Machine learning models that take computer program source code as input typically use Natural Language Processing (NLP) techniques. However, a major challenge is that code is written using an open, rapidly changing vocabulary due to, e.g., the coinage of new variable and method names. Reasoning over such a vocabulary is not something for which most NLP methods are designed. We introduce a Graph-Structured Cache to address this problem; this cache contains a node for each new word the model encounters with edges connecting each word to its occurrences in the code. We find that combining this graph-structured cache strategy with recent Graph-Neural-Network-based models for supervised learning on code improves the models' performance on a code completion task and a variable naming task --- with over 100\% relative improvement on the latter --- at the cost of a moderate increase in computation time.

\*\*\*\*\*

## BNN+: Improved Binary Network Training

Sajad Darabi,Mouloud Belbahri,Matthieu Courbariaux,Vahid Partovi Nia

Deep neural networks (DNN) are widely used in many applications. However, their deployment on edge devices has been difficult because they are resource hungry. Binary neural networks (BNN) help to alleviate the prohibitive resource requirements of DNN, where both activations and weights are limited to 1-bit. We propose an improved binary training method (BNN+), by introducing a regularization function that encourages training weights around binary values. In addition to this, to enhance model performance we add trainable scaling factors to our regularization functions. Furthermore, we use an improved approximation of the derivative of the sign activation function in the backward computation. These additions are based on linear operations that are easily implementable into the binary training framework. We show experimental results on CIFAR-10 obtaining an accuracy of 86.5%, on AlexNet and 91.3% with VGG network. On ImageNet, our method also outperforms the traditional BNN method and XNOR-net, using AlexNet by a margin of 4% and 2% top-1 accuracy respectively.

\*\*\*\*\*

## 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

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.

\*\*\*\*\*

## DEFactor: Differentiable Edge Factorization-based Probabilistic Graph Generation

Rim Assouel,Mohamed Ahmed,Marwin Segler,Amir Saffari,Yoshua Bengio

Generating novel molecules with optimal properties is a crucial step in many industries such as drug discovery.

Recently, deep generative models have shown a promising way of performing de-novo molecular design.

Although graph generative models are currently available they either have a graph size dependency in their number of parameters, limiting their use to only very small graphs or are formulated as a sequence of discrete actions needed to construct a graph, making the output graph non-differentiable w.r.t the model parameters, therefore preventing them to be used in scenarios such as conditional graph generation. In this work we propose a model for conditional graph generation t

that is computationally efficient and enables direct optimisation of the graph. We demonstrate favourable performance of our model on prototype-based molecular graph conditional generation tasks.

\*\*\*\*\*

#### MEAN-FIELD ANALYSIS OF BATCH NORMALIZATION

Mingwei Wei, James Stokes, David J Schwab

Batch Normalization (BatchNorm) is an extremely useful component of modern neural network architectures, enabling optimization using higher learning rates and achieving faster convergence. In this paper, we use mean-field theory to analytically quantify the impact of BatchNorm on the geometry of the loss landscape for multi-layer networks consisting of fully-connected and convolutional layers. We show that it has a flattening effect on the loss landscape, as quantified by the maximum eigenvalue of the Fisher Information Matrix. These findings are then used to justify the use of larger learning rates for networks that use BatchNorm, and we provide quantitative characterization of the maximal allowable learning rate to ensure convergence. Experiments support our theoretically predicted maximum learning rate, and furthermore suggest that networks with smaller values of the BatchNorm parameter achieve lower loss after the same number of epochs of training.

\*\*\*\*\*

End-to-end learning of pharmacological assays from high-resolution microscopy images

Markus Hofmarcher, Elisabeth Rumetshofer, Sepp Hochreiter, Günter Klambauer

Predicting the outcome of pharmacological assays based on high-resolution microscopy

images of treated cells is a crucial task in drug discovery which tremendously increases discovery rates. However, end-to-end learning on these images with convolutional neural networks (CNNs) has not been ventured for this task because it has been considered infeasible and overly complex. On the largest available public dataset, we compare several state-of-the-art CNNs trained in an end-to-end fashion with models based on a cell-centric approach involving segmentation.

We found that CNNs operating on full images containing hundreds of cells perform significantly better at assay prediction than networks operating

on a single-cell level. Surprisingly, we could predict 29% of the 209 pharmacological

assays at high predictive performance ( $AUC > 0.9$ ). We compared a novel CNN architecture called "GapNet" against four competing CNN architectures and found that it performs on par with the best methods and at the same time has the lowest training time. Our results demonstrate that end-to-end learning on

high-resolution imaging data is not only possible but even outperforms cell-centric

and segmentation-dependent approaches. Hence, the costly cell segmentation and feature extraction steps are not necessary, in fact they even hamper predictive performance.

Our work further suggests that many pharmacological assays could be replaced by high-resolution microscopy imaging together with convolutional neural networks.

\*\*\*\*\*

#### PCNN: Environment Adaptive Model Without Finetuning

Boyuan Feng, Kun Wan, Shu Yang, Yufei Ding

Convolutional Neural Networks (CNNs) have achieved tremendous success for many computer vision tasks, which shows a promising perspective of deploying CNNs on mobile platforms. An obstacle to this promising perspective is the tension between intensive resource consumption of CNNs and limited resource budget on mobile platforms. Existing works generally utilize a simpler architecture with lower accuracy for a higher energy-efficiency, \textit{i.e.}, trading accuracy for resource consumption. An emerging opportunity to both increasing accuracy and decreasing

ng resource consumption is **class skew**, *i.e.*, the strong temporal and spatial locality of the appearance of classes. However, it is challenging to efficiently utilize the class skew due to both the frequent switches and the huge number of class skews. Existing works use transfer learning to adapt the model towards the class skew during runtime, which consumes resource intensively. In this paper, we propose **probability layer**, an *easily-implemented and highly flexible add-on module* to adapt the model efficiently during runtime *without any fine-tuning* and achieving an *equivalent or better* performance than transfer learning. Further, both *increasing accuracy* and *decreasing resource consumption* can be achieved during runtime through the combination of probability layer and pruning methods.

\*\*\*\*\*

#### A More Globally Accurate Dimensionality Reduction Method Using Triplets

Ehsan Amid, Manfred K. Warmuth

We first show that the commonly used dimensionality reduction (DR) methods such as t-SNE and LargeVis poorly capture the global structure of the data in the low dimensional embedding. We show this via a number of tests for the DR methods that can be easily applied by any practitioner to the dataset at hand. Surprisingly enough, t-SNE performs the best w.r.t. the commonly used measures that reward the local neighborhood accuracy such as precision-recall while having the worst performance in our tests for global structure. We then contrast the performance of these two DR methods against our new method called TriMap. The main idea behind TriMap is to capture higher orders of structure with triplet information (instead of pairwise information used by t-SNE and LargeVis), and to minimize a robust loss function for satisfying the chosen triplets. We provide compelling experimental evidence on large natural datasets for the clear advantage of the TriMap DR results. As LargeVis, TriMap is fast and provides comparable runtime on large datasets.

\*\*\*\*\*

#### Temporal Difference Variational Auto-Encoder

Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, Theophane Weber

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

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

ll 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

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

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

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).

\*\*\*\*\*

Actor-Attention-Critic for Multi-Agent Reinforcement Learning

Shariq Iqbal, Fei Sha

Reinforcement learning in multi-agent scenarios is important for real-world appl



ications but presents challenges beyond those seen in single-agent settings. We present an actor-critic algorithm that trains decentralized policies in multi-agent settings, using centrally computed critics that share an attention mechanism which selects relevant information for each agent at every timestep. This attention mechanism enables more effective and scalable learning in complex multi-agent environments, when compared to recent approaches. Our approach is applicable not only to cooperative settings with shared rewards, but also individualized reward settings, including adversarial settings, and it makes no assumptions about the action spaces of the agents. As such, it is flexible enough to be applied to most multi-agent learning problems

\*\*\*\*\*

Overlapping Community Detection with Graph Neural Networks

Oleksandr Shchur, Stephan Günnemann

Community detection in graphs is of central importance in graph mining, machine learning and network science. Detecting overlapping communities is especially challenging, and remains an open problem. Motivated by the success of graph-based deep learning in other graph-related tasks, we study the applicability of this framework for overlapping community detection. We propose a probabilistic model for overlapping community detection based on the graph neural network architecture. Despite its simplicity, our model outperforms the existing approaches in the community recovery task by a large margin. Moreover, due to the inductive formulation, the proposed model is able to perform out-of-sample community detection for nodes that were not present at training time

\*\*\*\*\*

NLProlog: Reasoning with Weak Unification for Natural Language Question Answering

Leon Weber, Pasquale Minervini, Ulf Leser, Tim Rocktäschel

Symbolic logic allows practitioners to build systems that perform rule-based reasoning which is interpretable and which can easily be augmented with prior knowledge. However, such systems are traditionally difficult to apply to problems involving natural language due to the large linguistic variability of language. Currently, most work in natural language processing focuses on neural networks which learn distributed representations of words and their composition, thereby performing well in the presence of large linguistic variability. We propose to reap the benefits of both approaches by applying a combination of neural networks and logic programming to natural language question answering. We propose to employ an external, non-differentiable Prolog prover which utilizes a similarity function over pretrained sentence encoders. We fine-tune these representations via Evolution Strategies with the goal of multi-hop reasoning on natural language. This allows us to create a system that can apply rule-based reasoning to natural language and induce domain-specific natural language rules from training data. We evaluate the proposed system on two different question answering tasks, showing that it complements two very strong baselines – BIDAf (Seo et al., 2016a) and FASTQA (Weissenborn et al., 2017) – and outperforms both when used in an ensemble.

\*\*\*\*\*

Graph2Seq: Graph to Sequence Learning with Attention-Based Neural Networks

Kun Xu, Lingfei Wu, Zhiguo Wang, Yansong Feng, Michael Witbrock, Vadim Sheinin

The celebrated Sequence to Sequence learning (Seq2Seq) technique and its numerous variants achieve excellent performance on many tasks. However, many machine learning tasks have inputs naturally represented as graphs; existing Seq2Seq models face a significant challenge in achieving accurate conversion from graph form to the appropriate sequence. To address this challenge, we introduce a general end-to-end graph-to-sequence neural encoder-decoder architecture that maps an input graph to a sequence of vectors and uses an attention-based LSTM method to decode the target sequence from these vectors. Our method first generates the node and graph embeddings using an improved graph-based neural network with a novel aggregation strategy to incorporate edge direction information in the node embeddings. We further introduce an attention mechanism that aligns node embeddings and the decoding sequence to better cope with large graphs. Experimental results on bAbI, Shortest Path, and Natural Language Generation tasks demonstrate that our

r model achieves state-of-the-art performance and significantly outperforms existing graph neural networks, Seq2Seq, and Tree2Seq models; using the proposed bi-directional node embedding aggregation strategy, the model can converge rapidly to the optimal performance.

\*\*\*\*\*

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

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.

\*\*\*\*\*

Structured Prediction using cGANs with Fusion Discriminator

Faisal Mahmood, Wenhao Xu, Nicholas J. Durr, Jeremiah W. Johnson, Alan Yuille

We propose a novel method for incorporating conditional information into a generative adversarial network (GAN) for structured prediction tasks. This method is based on fusing features from the generated and conditional information in feature space and allows the discriminator to better capture higher-order statistics from the data. This method also increases the strength of the signals passed through the network where the real or generated data and the conditional data agree. The proposed method is conceptually simpler than the joint convolutional neural network - conditional Markov random field (CNN-CRF) models and enforces higher-order consistency without being limited to a very specific class of high-order potentials. Experimental results demonstrate that this method leads to improvement on a variety of different structured prediction tasks including image synthesis, semantic segmentation, and depth estimation.

\*\*\*\*\*

Constraining Action Sequences with Formal Languages for Deep Reinforcement Learning

Dong Xu, Eleanor Quint, Zeynep Hakkuguder, Haluk Dogan, Stephen Scott, Matthew Dwyer

We study the problem of deep reinforcement learning where the agent's action sequences are constrained, e.g., prohibition of dithering or overactuating action sequences that might damage a robot, drone, or other physical device. Our model focuses on constraints that can be described by automata such as DFAs or PDAs. We then propose multiple approaches to augment the state descriptions of the Markov decision process (MDP) with summaries of recent action histories. We empirically evaluate these methods applying DQN to three Atari games, training with reward shaping. We found that our approaches are effective in significantly reducing, and even eliminating, constraint violations while maintaining high reward. We also observed that the total reward achieved by an agent can be highly sensitive to how much the constraints encourage or discourage exploration of potentially effective actions during training, and, in addition to helping ensure safe policies, the use of constraints can enhance exploration during training.

\*\*\*\*\*

End-to-End Hierarchical Text Classification with Label Assignment Policy

Yuning Mao, Jingjing Tian, Jiawei Han, Xiang Ren

We present an end-to-end reinforcement learning approach to hierarchical text classification where documents are labeled by placing them at the right positions

in a given hierarchy.

While existing "global" methods construct hierarchical losses for model training, they either make "local" decisions at each hierarchy node or ignore the hierarchy structure during inference. To close the gap between training/inference and optimize holistic metrics in an end-to-end manner, we propose to learn a label assignment policy to determine where to place the documents and when to stop. The proposed method, HiLAP, optimizes holistic metrics over the hierarchy, makes inter-dependent decisions during inference, and can be combined with different text encoding models for end-to-end training.

Experiments on three public datasets show that HiLAP yields an average improvement of 33.4% in Macro-F1 and 5.0% in Samples-F1, outperforming state-of-the-art methods by a large margin.

\*\*\*\*\*

Learning Partially Observed PDE Dynamics with Neural Networks

Ibrahim Ayed, Emmanuel De Bézenac, Arthur Pajot, Patrick Gallinari

Spatio-Temporal processes bear a central importance in many applied scientific fields. Generally, differential equations are used to describe these processes. In this work, we address the problem of learning spatio-temporal dynamics with neural networks when only partial information on the system's state is available. Taking inspiration from the dynamical system approach, we outline a general framework in which complex dynamics generated by families of differential equations can be learned in a principled way. Two models are derived from this framework.

We demonstrate how they can be applied in practice by considering the problem of forecasting fluid flows. We show how the underlying equations fit into our formalism and evaluate our method by comparing with standard baselines.

\*\*\*\*\*

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

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, Jiajun Wu

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

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.

\*\*\*\*\*

## Deep Probabilistic Video Compression

Jun Han, Salvator Lombardo, Christopher Schroers, Stephan Mandt

We propose a variational inference approach to deep probabilistic video compression. Our model uses advances in variational autoencoders (VAEs) for sequential data and combines it with recent work on neural image compression. The approach jointly learns to transform the original video into a lower-dimensional representation as well as to entropy code this representation according to a temporally-conditioned probabilistic model. We split the latent space into local (per frame) and global (per segment) variables, and show that training the VAE to utilize both representations leads to an improved rate-distortion performance. Evaluation on small videos from public data sets with varying complexity and diversity shows that our model yields competitive results when trained on generic video content. Extreme compression performance is achieved for videos with specialized content if the model is trained on similar videos.

\*\*\*\*\*

## Reinforcement Learning with Perturbed Rewards

Jingkang Wang, Yang Liu, Bo Li

Recent studies have shown the vulnerability of reinforcement learning (RL) models in noisy settings. The sources of noises differ across scenarios. For instance, in practice, the observed reward channel is often subject to noise (e.g., when observed rewards are collected through sensors), and thus observed rewards may not be credible as a result. Also, in applications such as robotics, a deep reinforcement learning (DRL) algorithm can be manipulated to produce arbitrary errors. In this paper, we consider noisy RL problems where observed rewards by RL agents are generated with a reward confusion matrix. We call such observed rewards as perturbed rewards. We develop an unbiased reward estimator aided robust RL framework that enables RL agents to learn in noisy environments while observing only perturbed rewards. Our framework draws upon approaches for supervised learning with noisy data. The core ideas of our solution include estimating a reward confusion matrix and defining a set of unbiased surrogate rewards. We prove the convergence and sample complexity of our approach. Extensive experiments on different DRL platforms show that policies based on our estimated surrogate reward can achieve higher expected rewards, and converge faster than existing baselines. For instance, the state-of-the-art PPO algorithm is able to obtain 67.5% and 46.7% improvements in average on five Atari games, when the error rates are 10% and 30% respectively.

\*\*\*\*\*

## Evaluating GANs via Duality

Paulina Grnarova, Kfir Y Levy, Aurelien Lucchi, Nathanael Perraudin, Thomas Hofmann, Andreas Krause

Generative Adversarial Networks (GANs) have shown great results in accurately modeling complex distributions, but their training is known to be difficult due to instabilities caused by a challenging minimax optimization problem. This is especially troublesome given the lack of an evaluation metric that can reliably detect non-convergent behaviors. We leverage the notion of duality gap from game theory in order to propose a novel convergence metric for GANs that has low computational cost. We verify the validity of the proposed metric for various test scenarios commonly used in the literature.

\*\*\*\*\*

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

Xiangyi Chen, Sijia Liu, Ruoyu Sun, Mingyi Hong

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  $\mathcal{O}(\log\{T\}/\sqrt{T})$ .

{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.

\*\*\*\*\*

#### Controlling Over-generalization and its Effect on Adversarial Examples Detection and Generation

Mahdiah Abbasi, Arezoo Rajabi, Azadeh Sadat Mozafari, Rakesh B. Bobba, Christian Gagné

Convolutional Neural Networks (CNNs) significantly improve the state-of-the-art for many applications, especially in computer vision. However, CNNs still suffer from a tendency to confidently classify out-distribution samples from unknown classes into pre-defined known classes. Further, they are also vulnerable to adversarial examples. We are relating these two issues through the tendency of CNNs to over-generalize for areas of the input space not covered well by the training set. We show that a CNN augmented with an extra output class can act as a simple yet effective end-to-end model for controlling over-generalization. As an appropriate training set for the extra class, we introduce two resources that are computationally efficient to obtain: a representative natural out-distribution set and interpolated in-distribution samples. To help select a representative natural out-distribution set among available ones, we propose a simple measurement to assess an out-distribution set's fitness. We also demonstrate that training such an augmented CNN with representative out-distribution natural datasets and some interpolated samples allows it to better handle a wide range of unseen out-distribution samples and black-box adversarial examples without training it on any adversaries. Finally, we show that generation of white-box adversarial attacks using our proposed augmented CNN can become harder, as the attack algorithms have to get around the rejection regions when generating actual adversaries.

\*\*\*\*\*

#### Coverage and Quality Driven Training of Generative Image Models

Thomas LUCAS, Konstantin SHMELKOV, Karteek ALAHARI, Cordelia SCHMID, Jakob VERBEEK

Generative modeling of natural images has been extensively studied in recent years, yielding remarkable progress. Current state-of-the-art methods are either based on maximum likelihood estimation or adversarial training. Both methods have their own drawbacks, which are complementary in nature. The first leads to over-generalization as the maximum likelihood criterion encourages models to cover the support of the training data by heavily penalizing small masses assigned to training data. Simplifying assumptions in such models limits their capacity and makes them spill mass on unrealistic samples. The second leads to mode-dropping since adversarial training encourages high quality samples from the model, but only indirectly enforces diversity among the samples. To overcome these drawbacks we make two contributions. First, we propose a model that extends variational autoencoders by using deterministic invertible transformation layers to map samples from the decoder to the image space. This induces correlations among the pixels given the latent variables, improving over factorial decoders commonly used in variational autoencoders. Second, we propose a unified training approach that leverages coverage and quality based criteria. Our models obtain likelihood scores competitive with state-of-the-art likelihood-based models, while achieving sample quality typical of adversarially trained networks.

\*\*\*\*\*

#### Iteratively Learning from the Best

Yanyao Shen, Sujay Sanghavi

We study a simple generic framework to address the issue of bad training data; both bad labels in supervised problems, and bad samples in unsupervised ones. Our approach starts by fitting a model to the whole training dataset, but then iteratively improves it by alternating between (a) revisiting the training data to select samples with lowest current loss, and (b) re-training the model on only th

ese selected samples. It can be applied to any existing model training setting which provides a loss measure for samples, and a way to refit on new ones. We show the merit of this approach in both theory and practice. We first prove statistical consistency, and linear convergence to the ground truth and global optimum, for two simpler model settings: mixed linear regression, and gaussian mixture models. We then demonstrate its success empirically in (a) saving the accuracy of existing deep image classifiers when there are errors in the labels of training images, and (b) improving the quality of samples generated by existing DC-GAN models, when it is given training data that contains a fraction of the images from a different and unintended dataset. The experimental results show significant improvement over the baseline methods that ignore the existence of bad labels/samples.

\*\*\*\*\*

Gradient descent aligns the layers of deep linear networks

Ziwei Ji, Matus Telgarsky

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  $u_{iv_i} u_i^T$ ;
- (iii) these rank-1 matrices are aligned across layers, meaning  $|v_{i+1}^T u_i| \rightarrow 1$ .

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.

\*\*\*\*\*

Distinguishability of Adversarial Examples

Yi Qin, Ryan Hunt, Chuan Yue

Machine learning models including traditional models and neural networks can be easily fooled by adversarial examples which are generated from the natural examples with small perturbations. This poses a critical challenge to machine learning security, and impedes the wide application of machine learning in many important domains such as computer vision and malware detection. Unfortunately, even state-of-the-art defense approaches such as adversarial training and defensive distillation still suffer from major limitations and can be circumvented. From a unique angle, we propose to investigate two important research questions in this paper: Are adversarial examples distinguishable from natural examples? Are adversarial examples generated by different methods distinguishable from each other? These two questions concern the distinguishability of adversarial examples.

Answering them will potentially lead to a simple yet effective approach, termed as defensive distinction in this paper under the formulation of multi-label classification, for protecting against adversarial examples. We design and perform experiments using the MNIST dataset to investigate these two questions, and obtain highly positive results demonstrating the strong distinguishability of adversarial examples. We recommend that this unique defensive distinction approach should be seriously considered to complement other defense approaches.

\*\*\*\*\*

Hierarchical RL Using an Ensemble of Proprioceptive Periodic Policies

Kenneth Marino, Abhinav Gupta, Rob Fergus, Arthur Szlam

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 Schuster, Manmohan Chandraker

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

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

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.

\*\*\*\*\*

#### Prototypical Examples in Deep Learning: Metrics, Characteristics, and Utility

Nicholas Carlini, Ulfar Erlingsson, Nicolas Papernot

Machine learning (ML) research has investigated prototypes: examples that are representative of the behavior to be learned. We systematically evaluate five methods for identifying prototypes, both ones previously introduced as well as new ones we propose, finding all of them to provide meaningful but different interpretations. Through a human study, we confirm that all five metrics are well matched

d to human intuition. Examining cases where the metrics disagree offers an informative perspective on the properties of data and algorithms used in learning, with implications for data-corpus construction, efficiency, adversarial robustness, interpretability, and other ML aspects. In particular, we confirm that the "train on hard" curriculum approach can improve accuracy on many datasets and tasks, but that it is strictly worse when there are many mislabeled or ambiguous examples.

\*\*\*\*\*

#### Empirical Bounds on Linear Regions of Deep Rectifier Networks

Thiago Serra, Srikumar Ramalingam

One form of characterizing the expressiveness of a piecewise linear neural network is by the number of linear regions, or pieces, of the function modeled. We have observed substantial progress in this topic through lower and upper bounds on the maximum number of linear regions and a counting procedure. However, these bounds only account for the dimensions of the network and the exact counting may take a prohibitive amount of time, therefore making it infeasible to benchmark the expressiveness of networks. In this work, we approximate the number of linear regions of specific rectifier networks with an algorithm for probabilistic lower bounds of mixed-integer linear sets. In addition, we present a tighter upper bound that leverages network coefficients. We test both on trained networks. The algorithm for probabilistic lower bounds is several orders of magnitude faster than exact counting and the values reach similar orders of magnitude, hence making our approach a viable method to compare the expressiveness of such networks. The refined upper bound is particularly stronger on networks with narrow layers.

\*\*\*\*\*

#### Distilled Agent DQN for Provable Adversarial Robustness

Matthew Mirman, Marc Fischer, Martin Vechev

As deep neural networks have become the state of the art for solving complex reinforcement learning tasks, susceptibility to perceptual adversarial examples have become a concern. The transferability of adversarial examples is known to enable attacks capable of tricking the agent into bad states. In this work we demonstrate a simple poisoning attack able to keep deep RL from learning, and into fooling it when trained with defense methods commonly used for classification tasks. We then propose an algorithm called DadQN, based on deep Q-networks, which enables the use of stronger defenses, including defenses enabling the first ever on-line robustness certification of a deep RL agent.

\*\*\*\*\*

#### Visual Semantic Navigation using Scene Priors

Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, Roozbeh Mottaghi

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.

\*\*\*\*\*

#### Graph Learning Network: A Structure Learning Algorithm

Darwin Danilo Saire Pilco, Adín Ramírez Rivera

Graph prediction methods that work closely with the structure of the data, e.g., graph generation, commonly ignore the content of its nodes. On the other hand, the solutions that consider the node's information, e.g., classification, ignore the structure of the whole. And some methods exist in between, e.g., link prediction, but predict the structure piece-wise instead of considering the graph as a whole. We hypothesize that by jointly predicting the structure of the graph and its nodes' features, we can improve both tasks. We propose the Graph Learning



Network (GLN), a simple yet effective process to learn node embeddings and structure prediction functions. Our model uses graph convolutions to propose expected node features, and predict the best structure based on them. We repeat these steps sequentially to enhance the prediction and the embeddings. In contrast to existing generation methods that rely only on the structure of the data, we use the feature on the nodes to predict better relations, similar to what link prediction methods do. However, we propose an holistic approach to process the whole graph for our predictions. Our experiments show that our method predicts consistent structures across a set of problems, while creating meaningful node embeddings.

\*\*\*\*\*

Teaching to Teach by Structured Dark Knowledge

Ziliang Chen, Keze Wang, Liang Lin

To educate hyper deep learners, \emph{Curriculum Learnings} (CLs) require either human heuristic participation or self-deciding the difficulties of training instances. These coaching manners are blind to the coherent structures among examples, categories, and tasks, which are pregnant with more knowledgeable curriculum-routed teachers. In this paper, we propose a general methodology \emph{Teaching to Teach} (T2T). T2T is facilitated by \emph{Structured Dark Knowledge} (SDK) that constitutes a communication protocol between structured knowledge prior and teaching strategies. On one hand, SDK adaptively extracts structured knowledge by selecting a training subset consistent with the previous teaching decisions. On the other hand, SDK teaches curriculum-agnostic teachers by transferring this knowledge to update their teaching policy. This virtuous cycle can be flexibly deployed in most existing CL platforms and more importantly, very generic across various structured knowledge characteristics, e.g., diversity, complementarity, and causality. We evaluate T2T across different learners, teachers, and tasks, which significantly demonstrates that structured knowledge can be inherited by the teachers to further benefit learners' training.

\*\*\*\*\*

Alignment Based Matching Networks for One-Shot Classification and Open-Set Recognition

Paresh Malalur, Tommi Jaakkola

Deep learning for object classification relies heavily on convolutional models. While effective, CNNs are rarely interpretable after the fact. An attention mechanism can be used to highlight the area of the image that the model focuses on thus offering a narrow view into the mechanism of classification. We expand on this idea by forcing the method to explicitly align images to be classified to reference images representing the classes. The mechanism of alignment is learned and therefore does not require that the reference objects are anything like those being classified. Beyond explanation, our exemplar based cross-alignment method enables classification with only a single example per category (one-shot). Our model cuts the 5-way, 1-shot error rate in Omniglot from 2.1\% to 1.4\% and in MiniImageNet from 53.5\% to 46.5\% while simultaneously providing point-wise alignment information providing some understanding on what the network is capturing. This method of alignment also enables the recognition of an unsupported class (open-set) in the one-shot setting while maintaining an F1-score of above 0.5 for Omniglot even with 19 other distracting classes while baselines completely fail to separate the open-set class in the one-shot setting.

\*\*\*\*\*

AdaShift: Decorrelation and Convergence of Adaptive Learning Rate Methods

Zhiming Zhou\*, Qingru Zhang\*, Guansong Lu, Hongwei Wang, Weinan Zhang, Yong Yu

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 like

ly 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.

\*\*\*\*\*

#### Empirical Study of Easy and Hard Examples in CNN Training

Ikki Kishida, Hideki Nakayama

Deep Neural Networks (DNNs) generalize well despite their massive size and capability of memorizing all examples.

There is a hypothesis that DNNs start learning from simple patterns based on the observations that are consistently well-classified at early epochs (i.e., easy examples) and examples misclassified (i.e., hard examples).

However, despite the importance of understanding the learning dynamics of DNNs, properties of easy and hard examples are not fully investigated.

In this paper, we study the similarities of easy and hard examples respectively among different CNNs, assessing those examples' contributions to generalization. Our results show that most easy examples are identical among different CNNs, as they share similar dataset-dependent patterns (e.g., colors, structures, and superficial cues in high-frequency).

Moreover, while hard examples tend to contribute more to generalization than easy examples, removing a large number of easy examples leads to poor generalization, and we find that most misclassified examples in validation dataset are hard examples.

By analyzing intriguing properties of easy and hard examples, we discover that the reason why easy and hard examples have such properties can be explained by biases in a dataset and Stochastic Gradient Descent (SGD).

\*\*\*\*\*

#### Sliced Wasserstein Auto-Encoders

Soheil Kolouri, Phillip E. Pope, Charles E. Martin, Gustavo K. Rohde

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.

\*\*\*\*\*

#### Unifying Bilateral Filtering and Adversarial Training for Robust Neural Networks

Neale Ratzlaff, Li Fuxin

Recent analysis of deep neural networks has revealed their vulnerability to carefully structured adversarial examples. Many effective algorithms exist to craft these adversarial examples, but performant defenses seem to be far away. In this work, we explore the use of edge-aware bilateral filtering as a projection back to the space of natural images. We show that bilateral filtering is an effective defense in multiple attack settings, where the strength of the adversary gradually increases. In the case of adversary who has no knowledge of the defense, bilateral filtering can remove more than 90% of adversarial examples from a variety of different attacks. To evaluate against an adversary with complete knowledge

e of our defense, we adapt the bilateral filter as a trainable layer in a neural network and show that adding this layer makes ImageNet images significantly more robust to attacks. When trained under a framework of adversarial training, we show that the resulting model is hard to fool with even the best attack methods.

\*\*\*\*\*

#### Amortized Bayesian Meta-Learning

Sachin Ravi, Alex Beatson

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

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.

\*\*\*\*\*

#### CNNSAT: Fast, Accurate Boolean Satisfiability using Convolutional Neural Networks

Yu Wang, Fengjuan Gao, Amin Alipour, Linzhang Wang, Xuandong Li, Zhendong Su

Boolean satisfiability (SAT) is one of the most well-known NP-complete problems and has been extensively studied. State-of-the-art solvers exist and have found a wide range of applications. However, they still do not scale well to formulas with hundreds of variables. To tackle this fundamental scalability challenge, we introduce CNNSAT, a fast and accurate statistical decision procedure for SAT based on convolutional neural networks. CNNSAT's effectiveness is due to a precise and compact representation of Boolean formulas. On both real and synthetic formulas, CNNSAT is highly accurate and orders of magnitude faster than the state-of-the-art solver Z3. We also describe how to extend CNNSAT to predict satisfying assignments when it predicts a formula to be satisfiable.

\*\*\*\*\*

## LARGE BATCH SIZE TRAINING OF NEURAL NETWORKS WITH ADVERSARIAL TRAINING AND SECOND-ORDER INFORMATION

Zhewei Yao, Amir Gholami, Kurt Keutzer, Michael Mahoney

Stochastic Gradient Descent (SGD) methods using randomly selected batches are widely-used to train neural network (NN) models. Performing design exploration to find the best NN for a particular task often requires extensive training with different models on a large dataset, which is very computationally expensive. The most straightforward method to accelerate this computation is to distribute the batch of SGD over multiple processors. However, large batch training often times leads to degradation in accuracy, poor generalization, and even poor robustness to adversarial attacks. Existing solutions for large batch training either do not work or require massive hyper-parameter tuning. To address this issue, we propose a novel large batch training method which combines recent results in adversarial training (to regularize against "sharp minima") and second order optimization (to use curvature information to change batch size adaptively during training). We extensively evaluate our method on Cifar-10/100, SVHN, TinyImageNet, and ImageNet datasets, using multiple NNs, including residual networks as well as compressed networks such as SqueezeNext. Our new approach exceeds the performance of the existing solutions in terms of both accuracy and the number of SGD iterations (up to 1% and 3 $\times$ , respectively). We emphasize that this is achieved without any additional hyper-parameter tuning to tailor our method to any of these experiments.

\*\*\*\*\*

## Learning Protein Structure with a Differentiable Simulator

John Ingraham, Adam Riesselman, Chris Sander, Debora Marks

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.

\*\*\*\*\*

## Metropolis-Hastings view on variational inference and adversarial training

Kirill Neklyudov, Dmitry Vetrov

In this paper we propose to view the acceptance rate of the Metropolis-Hastings algorithm as a universal objective for learning to sample from target distribution -- given either as a set of samples or in the form of unnormalized density. This point of view unifies the goals of such approaches as Markov Chain Monte Carlo (MCMC), Generative Adversarial Networks (GANs), variational inference. To reveal the connection we derive the lower bound on the acceptance rate and treat it as the objective for learning explicit and implicit samplers. The form of the lower bound allows for doubly stochastic gradient optimization in case the target distribution factorizes (i.e. over data points). We empirically validate our approach on Bayesian inference for neural networks and generative models for images.

\*\*\*\*\*

## Shallow Learning For Deep Networks

Eugene Belilovsky, Michael Eickenberg, Edouard Oyallon

Shallow supervised 1-hidden layer neural networks have a number of favorable pro

properties that make them easier to interpret, analyze, and optimize than their deep counterparts, but lack their representational power. Here we use 1-hidden layer learning problems to sequentially build deep networks layer by layer, which can inherit properties from shallow networks. Contrary to previous approaches using shallow networks, we focus on problems where deep learning is reported as critical for success. We thus study CNNs on image recognition tasks using the large-scale Imagenet dataset and the CIFAR-10 dataset. Using a simple set of ideas for architecture and training we find that solving sequential 1-hidden-layer auxiliary problems leads to a CNN that exceeds AlexNet performance on ImageNet. Extending our training methodology to construct individual layers by solving 2- and 3-hidden layer auxiliary problems, we obtain an 11-layer network that exceeds VGG-11 on ImageNet obtaining 89.8% top-5 single crop. To our knowledge, this is the first competitive alternative to end-to-end training of CNNs that can scale to ImageNet. We conduct a wide range of experiments to study the properties this induces on the intermediate layers.

\*\*\*\*\*

Where and when to look? Spatial-temporal attention for action recognition in videos

Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Frederick Tung, Leonid Sigal

Inspired by the observation that humans are able to process videos efficiently by only paying attention when and where it is needed, we propose a novel spatial-temporal attention mechanism for video-based action recognition. For spatial attention, we learn a saliency mask to allow the model to focus on the most salient parts of the feature maps.

For temporal attention, we employ a soft temporal attention mechanism to identify the most relevant frames from an input video. Further, we propose a set of regularizers that ensure that our attention mechanism attends to coherent regions in space and time. Our model is efficient, as it proposes a separable spatio-temporal mechanism for video attention, while being able to identify important parts of the video both spatially and temporally. We demonstrate the efficacy of our approach on three public video action recognition datasets. The proposed approach leads to state-of-the-art performance on all of them, including the new large-scale Moments in Time dataset. Furthermore, we quantitatively and qualitatively evaluate our model's ability to accurately localize discriminative regions spatially and critical frames temporally. This is despite our model only being trained with per video classification labels.

\*\*\*\*\*

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

Pankaj Gupta, Yatin Chaudhary, Florian Buettner, Hinrich Schuetze

We address two challenges of probabilistic topic modelling in order to better estimate

the probability of a word in a given context, i.e.,  $P(\text{word} | \text{context})$ : (1) No Language Structure in Context: Probabilistic topic models ignore word order by summarizing a given context as a "bag-of-words" 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 word-topic 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-the-art 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.

\*\*\*\*\*

Learning and Planning with a Semantic Model

Yi Wu, Yuxin Wu, Aviv Tamar, Stuart Russell, Georgia Gkioxari, Yuandong Tian

Building deep reinforcement learning agents that can generalize and adapt to unseen environments remains a fundamental challenge for AI. This paper describes progresses on this challenge in the context of man-made environments, which are visually diverse but contain intrinsic semantic regularities. We propose a hybrid model-based and model-free approach, LEARNING and Planning with Semantics (LEAPS), consisting of a multi-target sub-policy that acts on visual inputs, and a Bayesian model over semantic structures. When placed in an unseen environment, the agent plans with the semantic model to make high-level decisions, proposes the next sub-target for the sub-policy to execute, and updates the semantic model based on new observations. We perform experiments in visual navigation tasks using House3D, a 3D environment that contains diverse human-designed indoor scenes with real-world objects. LEAPS outperforms strong baselines that do not explicitly plan using the semantic content.

\*\*\*\*\*

Neural Program Repair by Jointly Learning to Localize and Repair

Marko Vasic, Aditya Kanade, Petros Maniatis, David Bieber, Rishabh Singh

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.

\*\*\*\*\*

Adversarial Examples Are a Natural Consequence of Test Error in Noise

Nicolas Ford, Justin Gilmer, Ekin D. Cubuk

Over the last few years, the phenomenon of adversarial examples --- maliciously constructed inputs that fool trained machine learning models --- has captured the attention of the research community, especially when the adversary is restricted to making small modifications of a correctly handled input. At the same time, less surprisingly, image classifiers lack human-level performance on randomly corrupted images, such as images with additive Gaussian noise. In this work, we show that these are two manifestations of the same underlying phenomenon. We establish this connection in several ways. First, we find that adversarial examples exist at the same distance scales we would expect from a linear model with the same performance on corrupted images. Next, we show that Gaussian data augmentation during training improves robustness to small adversarial perturbations and that adversarial training improves robustness to several types of image corrup

tions. Finally, we present a model-independent upper bound on the distance from a corrupted image to its nearest error given test performance and show that in practice we already come close to achieving the bound, so that improving robustness further for the corrupted image distribution requires significantly reducing test error. All of this suggests that improving adversarial robustness should go hand in hand with improving performance in the presence of more general and realistic image corruptions. This yields a computationally tractable evaluation metric for defenses to consider: test error in noisy image distributions.

\*\*\*\*\*

HAPPIER: Hierarchical Polyphonic Music Generative RNN

Tianyang Zhao, Xiaoxuan Ma, Honglin Ma, Yizhou Wang

Generating polyphonic music with coherent global structure is a major challenge for automatic composition algorithms. The primary difficulty arises due to the inefficiency of models to recognize underlying patterns beneath music notes across different levels of time scales and remain long-term consistency while composing. Hierarchical architectures can capture and represent learned patterns in different temporal scales and maintain consistency over long time spans, and this corresponds to the hierarchical structure in music. Motivated by this, focusing on leveraging the idea of hierarchical models and improve them to fit the sequence modeling problem, our paper proposes HAPPIER: a novel Hierarchical Polyphonic music generative RNN. In HAPPIER, A higher 'measure level' learns correlations across measures and patterns for chord progressions, and a lower 'note level' learns a conditional distribution over the notes to generate within a measure. The two hierarchies operate at different clock rates: the higher one operates on a longer timescale and updates every measure, while the lower one operates on a shorter timescale and updates every unit duration. The two levels communicate with each other, and thus the entire architecture is trained jointly end-to-end by back-propagation. HAPPIER, profited from the strength of the hierarchical structure, generates polyphonic music with long-term dependencies compared to the state-of-the-art methods.

\*\*\*\*\*

Improved Language Modeling by Decoding the Past

Siddhartha Brahma

Highly regularized LSTMs achieve impressive results on several benchmark datasets in language modeling. We propose a new regularization method based on decoding the last token in the context using the predicted distribution of the next token. This biases the model towards retaining more contextual information, in turn improving its ability to predict the next token. With negligible overhead in the number of parameters and training time, our Past Decode Regularization (PDR) method achieves a word level perplexity of 55.6 on the Penn Treebank and 63.5 on the WikiText-2 datasets using a single softmax. We also show gains by using PDR in combination with a mixture-of-softmaxes, achieving a word level perplexity of 53.8 and 60.5 on these datasets. In addition, our method achieves 1.169 bits-per-character on the Penn Treebank Character dataset for character level language modeling. These results constitute a new state-of-the-art in their respective settings.

\*\*\*\*\*

Rating Continuous Actions in Spatial Multi-Agent Problems

Uwe Dick, Maryam Tavakol, Ulf Brefeld

We study credit assignment problems in spatial multi-agent environments where agents pursue a joint objective. On the example of soccer, we rate the movements of individual players with respect to their potential for staging a successful attack. We propose a purely data-driven approach to simultaneously learn a model of agent movements as well as their ratings via an agent-centric deep reinforcement learning framework. Our model allows for efficient learning and sampling of ratings in the continuous action space. We empirically observe on historic soccer data that the model accurately rates agent movements w.r.t. their relative contribution to the collective goal.

\*\*\*\*\*

Human-level Protein Localization with Convolutional Neural Networks

Elisabeth Rumetshofer, Markus Hofmarcher, Clemens Röhrig, Sepp Hochreiter, Günter Klambauer

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 propose 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

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

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.

\*\*\*\*\*

Towards the Latent Transcriptome



Assya Trofimov, Francis Dutil, Claude Perreault, Sebastien Lemieux, Yoshua Bengio, Joseph Paul Cohen

In this work we propose a method to compute continuous embeddings for kmers from raw RNA-seq data, in a reference-free fashion. We report that our model captures information of both DNA sequence similarity as well as DNA sequence abundance in the embedding latent space. We confirm the quality of these vectors by comparing them to known gene sub-structures and report that the latent space recovers exon information from raw RNA-Seq data from acute myeloid leukemia patients. Furthermore we show that this latent space allows the detection of genomic abnormalities such as translocations as well as patient-specific mutations, making this representation space both useful for visualization as well as analysis.

\*\*\*\*\*

Generalized Capsule Networks with Trainable Routing Procedure

Zhenhua Chen, Chuhua Wang, Tiancong Zhao, David Crandall

CapsNet (Capsule Network) was first proposed by Sabour et al. (2017) and later another version of CapsNet was proposed by Hinton et al. (2018). CapsNet has been proved effective in modeling spatial features with much fewer parameters. However, the routing procedures (dynamic routing and EM routing) in both papers are not well incorporated into the whole training process, and the optimal number for the routing procedure has to be found manually. We propose Generalized CapsNet (G-CapsNet) to overcome this disadvantage by incorporating the routing procedure into the optimization. We implement two versions of G-CapsNet (fully-connected and convolutional) on CAFFE (Jia et al. (2014)) and evaluate them by testing the accuracy on MNIST & CIFAR10, the robustness to white-box & black-box attack, and the generalization ability on GAN-generated synthetic images. We also explore the scalability of G-CapsNet by constructing a relatively deep G-CapsNet. The experiment shows that G-CapsNet has good generalization ability and scalability.

\*\*\*\*\*

Variational Autoencoder with Arbitrary Conditioning

Oleg Ivanov, Michael Figurnov, Dmitry Vetrov

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.

\*\*\*\*\*

Selectivity metrics can overestimate the selectivity of units: a case study on AlexNet

Ella M. Gale, Anh Nguyen, Ryan Blything, Nicholas Martin and Jeffrey S. Bowers

Various methods of measuring unit selectivity have been developed in order to understand the representations learned by neural networks (NNs). Here we undertake a comparison of four such measures on AlexNet, namely, localist selectivity, precision (Zhou et al, ICLR 2015), class-conditional mean activity selectivity CCMAS (Morcos et al, ICLR 2018), and a new measure called top-class selectivity.

In contrast with previous work on recurrent neural networks (RNNs), we fail to find any 100% selective 'localist units' in AlexNet, and demonstrate that the precision and CCMAS measures provide a much higher level of selectivity than is warranted, with the most selective hidden units only responding strongly to a small minority of images from within a category. We also generated activation maximization (AM) images that maximally activated individual units and found that under 5% of units in fc6 and conv5 produced interpretable images of objects, whereas fc8 produced over 50% interpretable images. Furthermore, the interpretable images in the hidden layers were not associated with highly selective units. These findings highlight the problem with current selectivity measures and show that new measures are required in order to provide a better assessment of learned representations in NNs. We also consider why localist representations are learned in RNNs and not AlexNet.

\*\*\*\*\*

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

Ryan L. Murphy, Balasubramaniam Srinivasan, Vinayak Rao, Bruno Ribeiro

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

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 exploits 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

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 problems 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

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

r model to transfer learned behaviours across tasks.

\*\*\*\*\*

InstaGAN: Instance-aware Image-to-Image Translation

Sangwoo Mo, Minsu Cho, Jinwoo Shin

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

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

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.

\*\*\*\*\*

A Rotation and a Translation Suffice: Fooling CNNs with Simple Transformations

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, Aleksander Madry

We show that simple spatial transformations, namely translations and rotations alone, suffice to fool neural networks on a significant fraction of their inputs in multiple image classification tasks. Our results are in sharp contrast to previous work in adversarial robustness that relied on more complicated optimization approaches unlikely to appear outside a truly adversarial context. Moreover,

the misclassifying rotations and translations are easy to find and require only a few black-box queries to the target model. Overall, our findings emphasize the need to design robust classifiers even for natural input transformations in benign settings.

\*\*\*\*\*

Learning What to Remember: Long-term Episodic Memory Networks for Learning from Streaming Data

Hyunwoo Jung, Moonsu Han, Minki Kang, Sungju Hwang

Current generation of memory-augmented neural networks has limited scalability as they cannot efficiently process data that are too large to fit in the external memory storage. One example of this is lifelong learning scenario where the model receives unlimited length of data stream as an input which contains vast majority of uninformative entries. We tackle this problem by proposing a memory network fit for long-term lifelong learning scenario, which we refer to as Long-term Episodic Memory Networks (LEMN), that features a RNN-based retention agent that learns to replace less important memory entries based on the retention probability generated on each entry that is learned to identify data instances of generic importance relative to other memory entries, as well as its historical importance. Such learning of retention agent allows our long-term episodic memory network to retain memory entries of generic importance for a given task. We validate our model on a path-finding task as well as synthetic and real question answering tasks, on which our model achieves significant improvements over the memory augmented networks with rule-based memory scheduling as well as an RL-based baseline that does not consider relative or historical importance of the memory.

\*\*\*\*\*

Rethinking learning rate schedules for stochastic optimization

Rong Ge, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli

There is a stark disparity between the learning rate schedules used in the practice of large scale machine learning and what are considered admissible learning rate schedules prescribed in the theory of stochastic approximation. Recent results, such as in the 'super-convergence' methods which use oscillating learning rates, serve to emphasize this point even more.

One plausible explanation is that non-convex neural network training procedures are better suited to the use of fundamentally different learning rate schedules, such as the 'cut the learning rate every constant number of epochs' method (which more closely resembles an exponentially decaying learning rate schedule); note that this widely used schedule is in stark contrast to the polynomial decay schemes prescribed in the stochastic approximation literature, which are indeed shown to be (worst case) optimal for classes of convex optimization problems.

The main contribution of this work shows that the picture is far more nuanced, where we do not even need to move to non-convex optimization to show other learning rate schemes can be far more effective. In fact, even for the simple case of stochastic linear regression with a fixed time horizon, the rate achieved by any polynomial decay scheme is sub-optimal compared to the statistical minimax rate (by a factor of condition number); in contrast the 'cut the learning rate every constant number of epochs' provides an exponential improvement (depending only logarithmically on the condition number) compared to any polynomial decay scheme. Finally, it is important to ask if our theoretical insights are somehow fundamentally tied to quadratic loss minimization (where we have circumvented minimax lower bounds for more general convex optimization problems)? Here, we conjecture that recent results which make the gradient norm small at a near optimal rate, for both convex and non-convex optimization, may also provide more insights into learning rate schedules used in practice.

\*\*\*\*\*

Unsupervised Conditional Generation using noise engineered mode matching GAN

Deepak Mishra, Prathosh AP, Aravind J, Prashant Pandey, Santanu Chaudhury

Conditional generation refers to the process of sampling from an unknown distrib

ution conditioned on semantics of the data. This can be achieved by augmenting the generative model with the desired semantic labels, albeit it is not straightforward in an unsupervised setting where the semantic label of every data sample is unknown. In this paper, we address this issue by proposing a method that can generate samples conditioned on the properties of a latent distribution engineered in accordance with a certain data prior. In particular, a latent space inversion network is trained in tandem with a generative adversarial network such that the modal properties of the latent space distribution are induced in the data generating distribution. We demonstrate that our model, despite being fully unsupervised, is effective in learning meaningful representations through its mode matching property. We validate our method on multiple unsupervised tasks such as conditional generation, dataset attribute discovery and inference using three real world image datasets namely MNIST, CIFAR-10 and CELEB-A and show that the results are comparable to the state-of-the-art methods.

\*\*\*\*\*

InfoBot: Transfer and Exploration via the Information Bottleneck

Anirudh Goyal, Riashat Islam, DJ Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew B. Tsvinick, Yoshua Bengio, Sergey Levine

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  $\{ \text{decision states} \}$ . 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.

\*\*\*\*\*

Over-parameterization Improves Generalization in the XOR Detection Problem

Alon Brutzkus, Amir Globerson

Empirical evidence suggests that neural networks with ReLU activations generalize better with over-parameterization. However, there is currently no theoretical analysis that explains this observation. In this work, we study a simplified learning task with over-parameterized convolutional networks that empirically exhibit the same qualitative phenomenon. For this setting, we provide a theoretical analysis of the optimization and generalization performance of gradient descent. Specifically, we prove data-dependent sample complexity bounds which show that over-parameterization improves the generalization performance of gradient descent.

\*\*\*\*\*

Understand the dynamics of GANs via Primal-Dual Optimization

Songtao Lu, Rahul Singh, Xiangyi Chen, Yongxin Chen, Mingyi Hong

Generative adversarial network (GAN) is one of the best known unsupervised learning techniques these days due to its superior ability to learn data distributions. In spite of its great success in applications, GAN is known to be notoriously hard to train. The tremendous amount of time it takes to run the training algorithm and its sensitivity to hyper-parameter tuning have been haunting researchers in this area. To resolve these issues, we need to first understand how GANs work. Herein, we take a step toward this direction by examining the dynamics of GANs. We relate a large class of GANs including the Wasserstein GANs to max-min optimization problems with the coupling term being linear over the discriminator. By developing new primal-dual optimization tools, we show that, with a proper stepsize choice, the widely used first-order iterative algorithm in training GANs would in fact converge to a stationary solution with a sublinear rate. The same framework also applies to multi-task learning and distributional robust learning problems. We verify our analysis on numerical examples with both synthetic and

real data sets. We hope our analysis shed light on future studies on the theoretical properties of relevant machine learning problems.

\*\*\*\*\*

Learning Global Additive Explanations for Neural Nets Using Model Distillation

Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, Albert Gordo

Interpretability has largely focused on local explanations, i.e. explaining why a model made a particular prediction for a sample. These explanations are appealing due to their simplicity and local fidelity. However, they do not provide information about the general behavior of the model. We propose to leverage model distillation to learn global additive explanations that describe the relationship between input features and model predictions. These global explanations take the form of feature shapes, which are more expressive than feature attributions. Through careful experimentation, we show qualitatively and quantitatively that global additive explanations are able to describe model behavior and yield insights about models such as neural nets. A visualization of our approach applied to a neural net as it is trained is available at <https://youtu.be/ErQYwNqzEdc>

\*\*\*\*\*

An investigation of model-free planning

Arthur Guez, Mehdi Mirza, Karol Gregor, Rishabh Kabra, Sébastien Racanière, Théophane Weber, David Raposo, Adam Santoro, Laurent Orseau, Tom Eccles, Greg Wayne, David Silver, Timothy Lillicrap

The field of reinforcement learning (RL) is facing increasingly challenging domains with combinatorial complexity. For an RL agent to address these challenges, it is essential that it can plan effectively. Prior work has typically utilized an explicit model of the environment, combined with a specific planning algorithm (such as tree search). More recently, a new family of methods have been proposed that learn how to plan, by providing the structure for planning via an inductive bias in the function approximator (such as a tree structured neural network), trained end-to-end by a model-free RL algorithm. In this paper, we go even further, and demonstrate empirically that an entirely model-free approach, without special structure beyond standard neural network components such as convolutional networks and LSTMs, can learn to exhibit many of the hallmarks that we would typically associate with a model-based planner. We measure our agent's effectiveness at planning in terms of its ability to generalize across a combinatorial and irreversible state space, its data efficiency, and its ability to utilize additional thinking time. We find that our agent has the characteristics that one might expect to find in a planning algorithm. Furthermore, it exceeds the state-of-the-art in challenging combinatorial domains such as Sokoban and outperforms other model-free approaches that utilize strong inductive biases toward planning.

\*\*\*\*\*

HyperGAN: Exploring the Manifold of Neural Networks

Neale Ratzlaff, Li Fuxin

We introduce HyperGAN, a generative network that learns to generate all the weight parameters of deep neural networks. HyperGAN first transforms low dimensional noise into a latent space, which can be sampled from to obtain diverse, performant sets of parameters for a target architecture. We utilize an architecture that bears resemblance to generative adversarial networks, but we evaluate the likelihood of samples with a classification loss. This is equivalent to minimizing the KL-divergence between the generated network parameter distribution and an unknown true parameter distribution. We apply HyperGAN to classification, showing that HyperGAN can learn to generate parameters which solve the MNIST and CIFAR-10 datasets with competitive performance to fully supervised learning while learning a rich distribution of effective parameters. We also show that HyperGAN can also provide better uncertainty than standard ensembles. This is evaluated by the ability of HyperGAN-generated ensembles to detect out of distribution data as well as adversarial examples. We see that in addition to being highly accurate on inlier data, HyperGAN can provide reasonable uncertainty estimates.

\*\*\*\*\*

Stochastic Gradient Push for Distributed Deep Learning

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, Mike Rabbat

Large mini-batch parallel SGD is commonly used for distributed training of deep networks. Approaches that use tightly-coupled exact distributed averaging based on AllReduce are sensitive to slow nodes and high-latency communication. In this work we show the applicability of Stochastic Gradient Push (SGP) for distributed training. SGP uses a gossip algorithm called PushSum for approximate distributed averaging, allowing for much more loosely coupled communications which can be beneficial in high-latency or high-variability scenarios. The tradeoff is that approximate distributed averaging injects additional noise in the gradient which can affect the train and test accuracies. We prove that SGP converges to a stationary point of smooth, non-convex objective functions. Furthermore, we validate empirically the potential of SGP. For example, using 32 nodes with 8 GPUs per node to train ResNet-50 on ImageNet, where nodes communicate over 10Gbps Ethernet, SGP completes 90 epochs in around 1.5 hours while AllReduce SGD takes over 5 hours, and the top-1 validation accuracy of SGP remains within 1.2% of that obtained using AllReduce SGD.

\*\*\*\*\*

Multi-agent Deep Reinforcement Learning with Extremely Noisy Observations  
Ozsel Kilinc, Giovanni Montana

Multi-agent reinforcement learning systems aim to provide interacting agents with the ability to collaboratively learn and adapt to the behaviour of other agents. In many real-world applications, the agents can only acquire a partial view of the world. Here we consider a setting whereby most agents' observations are also extremely noisy, hence only weakly correlated to the true state of the environment. Under these circumstances, learning an optimal policy becomes particularly challenging, even in the unrealistic case that an agent's policy can be made conditional upon all other agents' observations. To overcome these difficulties, we propose a multi-agent deep deterministic policy gradient algorithm enhanced by a communication medium (MADDPG-M), which implements a two-level, concurrent learning mechanism. An agent's policy depends on its own private observations as well as those explicitly shared by others through a communication medium. At any given point in time, an agent must decide whether its private observations are sufficiently informative to be shared with others. However, our environments provide no explicit feedback informing an agent whether a communication action is beneficial, rather the communication policies must also be learned through experience concurrently to the main policies. Our experimental results demonstrate that the algorithm performs well in six highly non-stationary environments of progressively higher complexity, and offers substantial performance gains compared to the baselines.

\*\*\*\*\*

Seq2Slate: Re-ranking and Slate Optimization with RNNs

Irwan Bello, Sayali Kulkarni, Sagar Jain, Craig Boutilier, Ed Chi, Elad Eban, Xiyang Luo, Alan Mackey, Ofer Meshi

Ranking is a central task in machine learning and information retrieval. In this task, it is especially important to present the user with a slate of items that is appealing as a whole. This in turn requires taking into account interactions between items, since intuitively, placing an item on the slate affects the decision of which other items should be chosen alongside it.

In this work, we propose a sequence-to-sequence model for ranking called seq2slate. At each step, the model predicts the next item to place on the slate given the items already chosen. The recurrent nature of the model allows complex dependencies between items to be captured directly in a flexible and scalable way. We show how to learn the model end-to-end from weak supervision in the form of easily obtained click-through data. We further demonstrate the usefulness of our approach in experiments on standard ranking benchmarks as well as in a real-world recommendation system.

\*\*\*\*\*

Emerging Disentanglement in Auto-Encoder Based Unsupervised Image Content Transfer

Ori Press, Tomer Galanti, Sagie Benaim, Lior Wolf

We study the problem of learning to map, in an unsupervised way, between domains

$A$  and  $B$ , such that the samples  $\mathbf{b}$  in  $B$  contain all the information that exists in samples  $\mathbf{a}$  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  $\mathbf{a}$  from the first domain to the other domain, the missing information is replicated from an independent reference sample  $\mathbf{b}$  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

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\*

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

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 adve



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.

\*\*\*\*\*

#### Generating Realistic Stock Market Order Streams

Junyi Li, Xintong Wang, Yaoyang Lin, Arunesh Sinha, Michael P. Wellman

We propose an approach to generate realistic and high-fidelity stock market data based on generative adversarial networks.

We model the order stream as a stochastic process with finite history dependence, and employ a conditional Wasserstein GAN to capture history dependence of orders in a stock market.

We test our approach with actual market and synthetic data on a number of different statistics, and find the generated data to be close to real data.

\*\*\*\*\*

#### Learning State Representations in Complex Systems with Multimodal Data

Pavel Solovov, Vladimir Aliev, Pavel Ostyakov, Gleb Sterkin, Elizaveta Logacheva, Stan Troeshetov, Roman Suvorov, Anton Mashikhin, Oleg Khomenko, Sergey I. Nikolenko

Representation learning becomes especially important for complex systems with multimodal data sources such as cameras or sensors. Recent advances in reinforcement learning and optimal control make it possible to design control algorithms on these latent representations, but the field still lacks a large-scale standard dataset for unified comparison. In this work, we present a large-scale dataset and evaluation framework for representation learning for the complex task of landing an airplane. We implement and compare several approaches to representation learning on this dataset in terms of the quality of simple supervised learning tasks and disentanglement scores. The resulting representations can be used for further tasks such as anomaly detection, optimal control, model-based reinforcement learning, and other applications.

\*\*\*\*\*

#### Unsupervised Speech Recognition via Segmental Empirical Output Distribution Matching

Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, Dong Yu

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.

\*\*\*\*\*

#### Label Propagation Networks

Kojin Oshiba, Nir Rosenfeld, Amir Globerson

Graph networks have recently attracted considerable interest, and in particular in the context of semi-supervised learning. These methods typically work by generating node representations that are propagated throughout a given weighted graph

h.

Here we argue that for semi-supervised learning, it is more natural to consider propagating labels in the graph instead. Towards this end, we propose a differentiable neural version of the classic Label Propagation (LP) algorithm. This formulation can be used for learning edge weights, unlike other methods where weights are set heuristically. Starting from a layer implementing a single iteration of LP, we proceed by adding several important non-linear steps that significantly enhance the label-propagating mechanism.

Experiments in two distinct settings demonstrate the utility of our approach.

\*\*\*\*\*

#### Expressiveness in Deep Reinforcement Learning

Xufang Luo, Qi Meng, Di He, Wei Chen, Yunhong Wang, Tie-Yan Liu

Representation learning in reinforcement learning (RL) algorithms focuses on extracting useful features for choosing good actions. Expressive representations are essential for learning well-performed policies. In this paper, we study the relationship between the state representation assigned by the state extractor and the performance of the RL agent. We observe that representations assigned by the better state extractor are more scattered than which assigned by the worse one. Moreover, RL agents achieving high performances always have high rank matrices which are composed by their representations. Based on our observations, we formally define expressiveness of the state extractor as the rank of the matrix composed by representations. Therefore, we propose to promote expressiveness so as to improve algorithm performances, and we call it Expressiveness Promoted DRL. We apply our method on both policy gradient and value-based algorithms, and experimental results on 55 Atari games show the superiority of our proposed method.

\*\*\*\*\*

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

Amanpreet Singh, Tushar Jain, Sainbayar Sukhbaatar

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 for each 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.

\*\*\*\*\*

#### Improving Sentence Representations with Multi-view Frameworks

Shuai Tang, Virginia R. de Sa

Multi-view learning can provide self-supervision when different views are available of the same data. Distributional hypothesis provides another form of useful self-supervision from adjacent sentences which are plentiful in large unlabelled corpora. Motivated by the asymmetry in the two hemispheres of the human brain as well as the observation that different learning architectures tend to emphasize different aspects of sentence meaning, we present two multi-view frameworks for learning sentence representations in an unsupervised fashion. One framework uses a generative objective and the other a discriminative one. In both frameworks, the final representation is an ensemble of two views, in which, one view encodes the input sentence with a Recurrent Neural Network (RNN), and the other view encodes it with a simple linear model. We show that, after learning, the vectors produced by our multi-view frameworks provide improved representations over the

ir single-view learnt counterparts, and the combination of different views gives representational improvement over each view and demonstrates solid transferability on standard downstream tasks.

\*\*\*\*\*

#### ProxQuant: Quantized Neural Networks via Proximal Operators

Yu Bai, Yu-Xiang Wang, Edo Liberty

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.

\*\*\*\*\*

#### Relational Graph Attention Networks

Dan Busbridge, Dane Sherburn, Pietro Cavallo, Nils Y. Hammerla

We investigate Relational Graph Attention Networks, a class of models that extends non-relational graph attention mechanisms to incorporate relational information, opening up these methods to a wider variety of problems. A thorough evaluation of these models is performed, and comparisons are made against established benchmarks. To provide a meaningful comparison, we retrain Relational Graph Convolutional Networks, the spectral counterpart of Relational Graph Attention Networks, and evaluate them under the same conditions. We find that Relational Graph Attention Networks perform worse than anticipated, although some configurations are marginally beneficial for modelling molecular properties. We provide insights as to why this may be, and suggest both modifications to evaluation strategies, as well as directions to investigate for future work.

\*\*\*\*\*

#### Optimal Completion Distillation for Sequence Learning

Sara Sabour, William Chan, Mohammad Norouzi

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

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

ints 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 with in the same category - implying that it is easier 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

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

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.

\*\*\*\*\*

#### Out-of-Sample Extrapolation with Neuron Editing

Matthew Amodio, David van Dijk, Ruth Montgomery, Guy Wolf, Smita Krishnaswamy

While neural networks can be trained to map from one specific dataset to another, they usually do not learn a generalized transformation that can extrapolate accurately outside the space of training. For instance, a generative adversarial network (GAN) exclusively trained to transform images of cars from light to dark might not have the same effect on images of horses. This is because neural networks are good at generation within the manifold of the data that they are trained on. However, generating new samples outside of the manifold or extrapolating "out-of-sample" is a much harder problem that has been less well studied. To address this, we introduce a technique called neuron editing that learns how neurons encode an edit for a particular transformation in a latent space. We use an auto encoder to decompose the variation within the dataset into activations of different neurons and generate transformed data by defining an editing transformation on those neurons. By performing the transformation in a latent trained space, we encode fairly complex and non-linear transformations to the data with much simpler distribution shifts to the neuron's activations. We showcase our technique on image domain/style transfer and two biological applications: removal of batch artifacts representing unwanted noise and modeling the effect of drug treatments to predict synergy between drugs.

\*\*\*\*\*

Penetrating the Fog: the Path to Efficient CNN Models

Kun Wan, Boyuan Feng, Shu Yang, Yufei Ding

With the increasing demand to deploy convolutional neural networks (CNNs) on mobile platforms, the sparse kernel approach was proposed, which could save more parameters than the standard convolution while maintaining accuracy. However, despite the great potential, no prior research has pointed out how to craft an sparse kernel design with such potential (i.e., effective design), and all prior works just adopt simple combinations of existing sparse kernels such as group convolution. Meanwhile due to the large design space it is also impossible to try all combinations of existing sparse kernels. In this paper, we are the first in the field to consider how to craft an effective sparse kernel design by eliminating the large design space. Specifically, we present a sparse kernel scheme to illustrate how to reduce the space from three aspects. First, in terms of composition we remove designs composed of repeated layers. Second, to remove designs with large accuracy degradation, we find an unified property named~\emph{information field} behind various sparse kernel designs, which could directly indicate the final accuracy. Last, we remove designs in two cases where a better parameter efficiency could be achieved. Additionally, we provide detailed efficiency analysis on the final 4 designs in our scheme. Experimental results validate the idea of our scheme by showing that our scheme is able to find designs which are more efficient in using parameters and computation with similar or higher accuracy.

\*\*\*\*\*

Neural separation of observed and unobserved distributions

Tavi Halperin, Ariel Ephrat, Yedid Hoshen

Separating mixed distributions is a long standing challenge for machine learning and signal processing. Applications include: single-channel multi-speaker separation (cocktail party problem), singing voice separation and separating reflections from images. Most current methods either rely on making strong assumptions on the source distributions (e.g. sparsity, low rank, repetitiveness) or rely on having training samples of each source in the mixture. In this work, we tackle the scenario of extracting an unobserved distribution additively mixed with a signal from an observed (arbitrary) distribution. We introduce a new method: Neural Egg Separation - an iterative method that learns to separate the known distribution from progressively finer estimates of the unknown distribution. In some settings, Neural Egg Separation is initialization sensitive, we therefore introduce GLO Masking which ensures a good initialization. Extensive experiments show that our method outperforms current methods that use the same level of supervision and often achieves similar performance to full supervision.

\*\*\*\*\*

Beyond Games: Bringing Exploration to Robots in Real-world

Deepak Pathak, Dhiraj Gandhi, Abhinav Gupta

Exploration has been a long standing problem in both model-based and model-free learning methods for sensorimotor control. While there has been major advances over the years, most of these successes have been demonstrated in either video games or simulation environments. This is primarily because the rewards (even the intrinsic ones) are non-differentiable since they are function of the environment (which is a black-box). In this paper, we focus on the policy optimization aspect of the intrinsic reward function. Specifically, by using a local approximation, we formulate intrinsic reward as a differentiable function so as to perform policy optimization using likelihood maximization -- much like supervised learning instead of reinforcement learning. This leads to a significantly sample efficient exploration policy. Our experiments clearly show that our approach outperforms both on-policy and off-policy optimization approaches like REINFORCE and DQN respectively. But most importantly, we are able to implement an exploration policy on a robot which learns to interact with objects completely from scratch just using data collected via the differentiable exploration module. See project videos at <https://doubleblindICLR.github.io/robot-exploration/>

\*\*\*\*\*

FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models  
Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, David Duvenaud  
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.

\*\*\*\*\*

Feature quantization for parsimonious and interpretable predictive models  
Adrien EHRHARDT, Vincent VANDEWALLE, Christophe BIERNACKI, Philippe HEINRICH  
For regulatory and interpretability reasons, the logistic regression is still widely used by financial institutions to learn the refunding probability of a loan from applicant's historical data. To improve prediction accuracy and interpretability, a preprocessing step quantizing both continuous and categorical data is usually performed: continuous features are discretized by assigning factor levels to intervals and, if numerous, levels of categorical features are grouped. However, a better predictive accuracy can be reached by embedding this quantization estimation step directly into the predictive estimation step itself. By doing so, the predictive loss has to be optimized on a huge and untractable discontinuous quantization set. To overcome this difficulty, we introduce a specific two-step optimization strategy: first, the optimization problem is relaxed by approximating discontinuous quantization functions by smooth functions; second, the resulting relaxed optimization problem is solved via a particular neural network and stochastic gradient descent. The strategy gives then access to good candidates for the original optimization problem after a straightforward maximum a posteriori procedure to obtain cutpoints. The good performances of this approach, which we call *glm-disc*, are illustrated on simulated and real data from the UCI library and Cr dit Agricole Consumer Finance (a major European historic player in the consumer credit market). The results show that practitioners finally have an automatic all-in-one tool that answers their recurring needs of quantization for predictive tasks.

\*\*\*\*\*

Bias Also Matters: Bias Attribution for Deep Neural Network Explanation  
Shengjie Wang, Tianyi Zhou, Jeff Bilmes  
The gradient of a deep neural network (DNN) w.r.t. the input provides information that can be used to explain the output prediction in terms of the input features.

es and has been widely studied to assist in interpreting DNNs. In a linear model (i.e.,  $g(x)=wx+b$ ), the gradient corresponds solely to the weights  $w$ . Such a model can reasonably locally linearly approximate a smooth nonlinear DNN, and hence the weights of this local model are the gradient. The other part, however, of a local linear model, i.e., the bias  $b$ , is usually overlooked in attribution methods since it is not part of the gradient. In this paper, we observe that since the bias in a DNN also has a non-negligible contribution to the correctness of predictions, it can also play a significant role in understanding DNN behaviors. In particular, we study how to attribute a DNN's bias to its input features. We propose a backpropagation-type algorithm "bias back-propagation (BBp)" that starts at the output layer and iteratively attributes the bias of each layer to its input nodes as well as combining the resulting bias term of the previous layer. This process stops at the input layer, where summing up the attributions over all the input features exactly recovers  $b$ . Together with the backpropagation of the gradient generating  $w$ , we can fully recover the locally linear model  $g(x)=wx+b$ . Hence, the attribution of the DNN outputs to its inputs is decomposed into two parts, the gradient  $w$  and the bias attribution, providing separate and complementary explanations. We study several possible attribution methods applied to the bias of each layer in BBp. In experiments, we show that BBp can generate complementary and highly interpretable explanations of DNNs in addition to gradient-based attributions.

\*\*\*\*\*

RANDOM MASK: Towards Robust Convolutional Neural Networks

Tiangge Luo, Tianle Cai, Mengxiao Zhang, Siyu Chen, Liwei Wang

Robustness of neural networks has recently been highlighted by the adversarial examples, i.e., inputs added with well-designed perturbations which are imperceptible to humans but can cause the network to give incorrect outputs. In this paper, we design a new CNN architecture that by itself has good robustness. We introduce a simple but powerful technique, Random Mask, to modify existing CNN structures. We show that CNN with Random Mask achieves state-of-the-art performance against black-box adversarial attacks without applying any adversarial training. We next investigate the adversarial examples which "fool" a CNN with Random Mask. Surprisingly, we find that these adversarial examples often "fool" humans as well. This raises fundamental questions on how to define adversarial examples and robustness properly.

\*\*\*\*\*

Exploration by random network distillation

Yuri Burda, Harrison Edwards, Amos Storkey, Oleg Klimov

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 to 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.

\*\*\*\*\*

Deep Reinforcement Learning of Universal Policies with Diverse Environment Summaries

Felix Berkenkamp, Debadeepta Dey, Ashish Kapoor

Deep reinforcement learning has enabled robots to complete complex tasks in simulation. However, the resulting policies do not transfer to real robots due to model errors in the simulator. One solution is to randomize the simulation environment, so that the resulting, trained policy achieves high performance in expectation.

tion over a variety of configurations that could represent the real-world. However, the distribution over simulator configurations must be carefully selected to represent the relevant dynamic modes of the system, as otherwise it can be unlikely to sample challenging configurations frequently enough. Moreover, the ideal distribution to improve the policy changes as the policy (un)learns to solve tasks in certain configurations. In this paper, we propose to use an inexpensive, kernel-based summarization method that identifies configurations that lead to diverse behaviors. Since failure modes for the given task are naturally diverse, the policy trains on a mixture of representative and challenging configurations, which leads to more robust policies. In experiments, we show that the proposed method achieves the same performance as domain randomization in simple cases, but performs better when domain randomization does not lead to diverse dynamic modes.

\*\*\*\*\*

#### 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

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.

\*\*\*\*\*

#### Detecting Adversarial Examples Via Neural Fingerprinting

Sumanth Dathathri, Stephan Zheng, Yisong Yue, Richard M. Murray

Deep neural networks are vulnerable to adversarial examples: input data that has been manipulated to cause dramatic model output errors. To defend against such attacks, we propose NeuralFingerprinting: a simple, yet effective method to detect adversarial examples that verifies whether model behavior is consistent with a set of fingerprints. These fingerprints are encoded into the model response during training and are inspired by the use of biometric and cryptographic signatures. In contrast to previous defenses, our method does not rely on knowledge of the adversary and can scale to large networks and input data. The benefits of our method are that 1) it is fast, 2) it is prohibitively expensive for an attacker to reverse-engineer which fingerprints were used, and 3) it does not assume knowledge of the adversary. In this work, we 1) theoretically analyze NeuralFingerprinting for linear models and 2) show that NeuralFingerprinting significantly improves on state-of-the-art detection mechanisms for deep neural networks, by detecting the strongest known adversarial attacks with 98-100% AUC-ROC scores on the MNIST, CIFAR-10 and MiniImagenet (20 classes) datasets. In particular, we consider several threat models, including the most conservative one in which the attacker has full knowledge of the defender's strategy. In all settings, the detection accuracy of NeuralFingerprinting generalizes well to unseen test-data and is robust over a wide range of hyperparameters.

\*\*\*\*\*

#### Modeling Dynamics of Biological Systems with Deep Generative Neural Networks

Scott Gigante, David van Dijk, Kevin R. Moon, Alexander Strzalkowski, Katie Ferguson, Guy Wolf, Smita Krishnaswamy

Biological data often contains measurements of dynamic entities such as cells or organisms in various states of progression. However, biological systems are notoriously difficult to describe analytically due to their many interacting compon



ents, and in many cases, the technical challenge of taking longitudinal measurements. This leads to difficulties in studying the features of the dynamics, for example the drivers of the transition. To address this problem, we present a deep neural network framework we call Dynamics Modeling Network or DyMoN. DyMoN is a neural network framework trained as a deep generative Markov model whose next state is a probability distribution based on the current state. DyMoN is well-suited to the idiosyncrasies of biological data, including noise, sparsity, and the lack of longitudinal measurements in many types of systems. Thus, DyMoN can be trained using probability distributions derived from the data in any way, such as trajectories derived via dimensionality reduction methods, and does not require longitudinal measurements. We show the advantage of learning deep models over shallow models such as Kalman filters and hidden Markov models that do not learn representations of the data, both in terms of learning embeddings of the data and also in terms training efficiency, accuracy and ability to multitask. We perform three case studies of applying DyMoN to different types of biological systems and extracting features of the dynamics in each case by examining the learned model.

\*\*\*\*\*

Pseudosaccades: A simple ensemble scheme for improving classification performance of deep nets

Jin Sean Lim, Robert John Durrant

We describe a simple ensemble approach that, unlike conventional ensembles, uses multiple random data sketches ('pseudosaccades') rather than multiple classifiers

to improve classification performance. Using this simple, but novel, approach we obtain statistically significant improvements in classification performance on

AlexNet, GoogLeNet, ResNet-50 and ResNet-152 baselines on Imagenet data - e.g. of the order of 0.3% to 0.6% in Top-1 accuracy and similar improvements in Top-k accuracy - essentially nearly for free.

\*\*\*\*\*

Jumpout: Improved Dropout for Deep Neural Networks with Rectified Linear Units

Shengjie Wang, Tianyi Zhou, Jeff Bilmes

Dropout is a simple yet effective technique to improve generalization performance and prevent overfitting in deep neural networks (DNNs). In this paper, we discuss three novel observations about dropout to better understand the generalization of DNNs with rectified linear unit (ReLU) activations: 1) dropout is a smoothing technique that encourages each local linear model of a DNN to be trained on data points from nearby regions; 2) a constant dropout rate can result in effective neural-deactivation rates that are significantly different for layers with different fractions of activated neurons; and 3) the rescaling factor of dropout causes an inconsistency to occur between the normalization during training and testing conditions when batch normalization is also used. The above leads to three simple but nontrivial improvements to dropout resulting in our proposed method "Jumpout." Jumpout samples the dropout rate using a monotone decreasing distribution (such as the right part of a truncated Gaussian), so the local linear model at each data point is trained, with high probability, to work better for data points from nearby than from more distant regions. Instead of tuning a dropout rate for each layer and applying it to all samples, jumpout moreover adaptively normalizes the dropout rate at each layer and every training sample/batch, so the effective dropout rate applied to the activated neurons are kept the same. Moreover, we rescale the outputs of jumpout for a better trade-off that keeps both the variance and mean of neurons more consistent between training and test phases, which mitigates the incompatibility between dropout and batch normalization. Compared to the original dropout, jumpout shows significantly improved performance on CIFAR10, CIFAR100, Fashion-MNIST, STL10, SVHN, ImageNet-1k, etc., while introducing negligible additional memory and computation costs.

\*\*\*\*\*

Generative Question Answering: Learning to Answer the Whole Question

Mike Lewis, Angela Fan

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.

\*\*\*\*\*

Successor Uncertainties: exploration and uncertainty in temporal difference learning

David Janz, Jiri Hron, José Miguel Hernández-Lobato, Katja Hofmann, Sebastian Tschieschek

We consider the problem of balancing exploration and exploitation in sequential decision making problems. This trade-off naturally lends itself to probabilistic modelling. For a probabilistic approach to be effective, considering uncertainty about all immediate and long-term consequences of agent's actions is vital. An estimate of such uncertainty can be leveraged to guide exploration even in situations where the agent needs to perform a potentially long sequence of actions before reaching an under-explored area of the environment. This observation was made by the authors of the Uncertainty Bellman Equation model (O'Donoghue et al., 2018), which explicitly considers full marginal uncertainty for each decision the agent faces. However, their model still considers a fully factorised posterior over the consequences of each action, meaning that dependencies vital for correlated long-term exploration are ignored. We go a step beyond and develop Successor Uncertainties, a probabilistic model for the state-action value function of a Markov Decision Process with a non-factorised covariance. We demonstrate how this leads to greatly improved performance on classic tabular exploration benchmarks and show strong performance of our method on a subset of Atari baselines. Overall, Successor Uncertainties provides a better probabilistic model for temporal difference learning at a similar computational cost to its predecessors.

\*\*\*\*\*

Decoupled Weight Decay Regularization

Ilya Loshchilov, Frank Hutter

$L_2$  regularization and weight decay regularization are equivalent for standard stochastic gradient descent (when rescaled by the learning rate), but as we demonstrate this is *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 *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 <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

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.

\*\*\*\*\*

Detecting Out-Of-Distribution Samples Using Low-Order Deep Features Statistics  
Igor M. Quintanilha, Roberto de M. E. Filho, José Lezama, Mauricio Delbracio, Leonardo O. Nunes

The ability to detect when an input sample was not drawn from the training distribution is an important desirable property of deep neural networks. In this paper, we show that a simple ensembling of first and second order deep feature statistics can be exploited to effectively differentiate in-distribution and out-of-distribution samples. Specifically, we observe that the mean and standard deviation within feature maps differs greatly between in-distribution and out-of-distribution samples. Based on this observation, we propose a simple and efficient plug-and-play detection procedure that does not require re-training, pre-processing or changes to the model. The proposed method outperforms the state-of-the-art by a large margin in all standard benchmarking tasks, while being much simpler to implement and execute. Notably, our method improves the true negative rate from 39.6% to 95.3% when 95% of in-distribution (CIFAR-100) are correctly detected using a DenseNet and the out-of-distribution dataset is TinyImageNet resize. The source code of our method will be made publicly available.

\*\*\*\*\*

Learning Gibbs-regularized GANs with variational discriminator reparameterization

Nicholas Rhinehart, Anqi Liu, Kihyuk Sohn, Paul Vernaza

We propose a novel approach to regularizing generative adversarial networks (GANs) leveraging learned  $\{\text{structured Gibbs distributions}\}$ . Our method consists of reparameterizing the discriminator to be an explicit function of two densities: the generator PDF  $q$  and a structured Gibbs distribution  $\nu$ . Leveraging recent work on invertible pushforward density estimators, this reparameterization is made possible by assuming the generator is invertible, which enables the analytic evaluation of the generator PDF  $q$ . We further propose optimizing the Jeffrey divergence, which balances mode coverage with sample quality. The combination of this loss and reparameterization allows us to effectively regularize the generator by imposing structure from domain knowledge on  $\nu$ , as in classical graphical models. Applying our method to a vehicle trajectory forecasting task, we observe that we are able to obtain quantitatively superior mode coverage as well as better-quality samples compared to traditional methods.

\*\*\*\*\*

Learning to Represent Edits

Pengcheng Yin, Graham Neubig, Miltiadis Allamanis, Marc Brockschmidt, Alexander L. Gunt

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.

\*\*\*\*\*

Feature prioritization and regularization improve standard accuracy and adversarial robustness

Chihuang Liu, Joseph JaJa

Adversarial training has been successfully applied to build robust models at a certain cost. While the robustness of a model increases, the standard classification accuracy declines. This phenomenon is suggested to be an inherent trade-off.

We propose a model that employs feature prioritization by a nonlinear attention module and  $L_2$  feature regularization to improve the adversarial robustness and the standard accuracy relative to adversarial training. The attention module encourages the model to rely heavily on robust features by assigning larger weights to them while suppressing non-robust features. The regularizer encourages the model to extract similar features for the natural and adversarial images, effectively ignoring the added perturbation. In addition to evaluating the robustness of our model, we provide justification for the attention module and propose a novel experimental strategy that quantitatively demonstrates that our model is almost ideally aligned with salient data characteristics. Additional experimental results illustrate the power of our model relative to the state of the art methods.

\*\*\*\*\*

On the Computational Inefficiency of Large Batch Sizes for Stochastic Gradient Descent

Noah Golmant, Nikita Vemuri, Zhewei Yao, Vladimir Feinberg, Amir Gholami, Kai Rothaug, Michael Mahoney, Joseph Gonzalez

Increasing the mini-batch size for stochastic gradient descent offers significant opportunities to reduce wall-clock training time, but there are a variety of theoretical and systems challenges that impede the widespread success of this technique (Das et al., 2016; Keskar et al., 2016). We investigate these issues, with an emphasis on time to convergence and total computational cost, through an extensive empirical analysis of network training across several architectures and problem domains, including image classification, image segmentation, and language modeling. Although it is common practice to increase the batch size in order to fully exploit available computational resources, we find a substantially more nuanced picture. Our main finding is that across a wide range of network architectures and problem domains, increasing the batch size beyond a certain point yields no decrease in wall-clock time to convergence for either train or test loss.

This batch size is usually substantially below the capacity of current systems. We show that popular training strategies for large batch size optimization begin to fail before we can populate all available compute resources, and we show that the point at which these methods break down depends more on attributes like model architecture and data complexity than it does directly on the size of the dataset.

\*\*\*\*\*

Unsupervised Domain Adaptation for Distance Metric Learning

Kihyuk Sohn, Wenling Shang, Xiang Yu, Manmohan Chandraker

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 dom

ains 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

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

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.

\*\*\*\*\*

DEEP GEOMETRICAL GRAPH CLASSIFICATION

Mostafa Rahmani, Ping Li

Most of the existing Graph Neural Networks (GNNs) are the mere extension of the Convolutional Neural Networks (CNNs) to graphs. Generally, they consist of several steps of message passing between the nodes followed by a global indiscriminate feature pooling function. In many data-sets, however, the nodes are unlabeled or their labels provide no information about the similarity between the nodes and

d the locations of the nodes in the graph. Accordingly, message passing may not propagate helpful information throughout the graph. We show that this conventional approach can fail to learn to perform even simple graph classification tasks.

We alleviate this serious shortcoming of the GNNs by making them a two step method. In the first of the proposed approach, a graph embedding algorithm is utilized to obtain a continuous feature vector for each node of the graph. The embedding algorithm represents the graph as a point-cloud in the embedding space. In the second step, the GNN is applied to the point-cloud representation of the graph provided by the embedding method. The GNN learns to perform the given task by inferring the topological structure of the graph encoded in the spatial distribution of the embedded vectors. In addition, we extend the proposed approach to the graph clustering problem and a new architecture for graph clustering is proposed. Moreover, the spatial representation of the graph is utilized to design a graph pooling algorithm. We turn the problem of graph down-sampling into a column sampling problem, i.e., the sampling algorithm selects a subset of the nodes whose feature vectors preserve the spatial distribution of all the feature vectors.

We apply the proposed approach to several popular benchmark data-sets and it is shown that the proposed geometrical approach strongly improves the state-of-the-art result for several data-sets. For instance, for the PTC data-set, we improve the state-of-the-art result for more than 22 %.

\*\*\*\*\*

A comprehensive, application-oriented study of catastrophic forgetting in DNNs  
B. Pfülb, A. Gepperth

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

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.

\*\*\*\*\*

Normalization Gradients are Least-squares Residuals  
Yi Liu

Batch Normalization (BN) and its variants have seen widespread adoption in the deep learning community because they improve the training of deep neural networks. Discussions of why this normalization works so well remain unsettled. We make explicit the relationship between ordinary least squares and partial derivatives computed when back-propagating through BN. We recast the back-propagation of B

N as a least squares fit, which zero-centers and decorrelates partial derivatives from normalized activations. This view, which we term  $\{\text{gradient-least-squares}\}$ , is an extensible and arithmetically accurate description of BN. To further explore this perspective, we motivate, interpret, and evaluate two adjustments to BN.

\*\*\*\*\*

#### Convergent Reinforcement Learning with Function Approximation: A Bilevel Optimization Perspective

Zhuoran Yang, Zuyue Fu, Kaiqing Zhang, Zhaoran Wang

We study reinforcement learning algorithms with nonlinear function approximation in the online setting. By formulating both the problems of value function estimation and policy learning as bilevel optimization problems, we propose online Q-learning and actor-critic algorithms for these two problems respectively. Our algorithms are gradient-based methods and thus are computationally efficient. Moreover, by approximating the iterates using differential equations, we establish convergence guarantees for the proposed algorithms. Thorough numerical experiments are conducted to back up our theory.

\*\*\*\*\*

#### Inferring Reward Functions from Demonstrators with Unknown Biases

Rohin Shah, Noah Gundotra, Pieter Abbeel, Anca Dragan

Our goal is to infer reward functions from demonstrations. In order to infer the correct reward function, we must account for the systematic ways in which the demonstrator is suboptimal. Prior work in inverse reinforcement learning can account for specific, known biases, but cannot handle demonstrators with unknown biases. In this work, we explore the idea of learning the demonstrator's planning algorithm (including their unknown biases), along with their reward function. What makes this challenging is that any demonstration could be explained either by positing a term in the reward function, or by positing a particular systematic bias. We explore what assumptions are sufficient for avoiding this impossibility result: either access to tasks with known rewards which enable estimating the planner separately, or that the demonstrator is sufficiently close to optimal that this can serve as a regularizer. In our exploration with synthetic models of human biases, we find that it is possible to adapt to different biases and perform better than assuming a fixed model of the demonstrator, such as Boltzmann rationality.

\*\*\*\*\*

#### LEARNING GENERATIVE MODELS FOR DEMIXING OF STRUCTURED SIGNALS FROM THEIR SUPERPOSITION USING GANS

Mohammadreza Soltani, Swayambhoo Jain, Abhinav V. Sivasubramanian

Recently, Generative Adversarial Networks (GANs) have emerged as a popular alternative for modeling complex high dimensional distributions. Most of the existing works implicitly assume that the clean samples from the target distribution are easily available. However, in many applications, this assumption is violated. In this paper, we consider the problem of learning GANs under the observation setting when the samples from target distribution are given by the superposition of two structured components. We propose two novel frameworks: denoising-GAN and demixing-GAN. The denoising-GAN assumes access to clean samples from the second component and try to learn the other distribution, whereas demixing-GAN learns the distribution of the components at the same time. Through comprehensive numerical experiments, we demonstrate that proposed frameworks can generate clean samples from unknown distributions, and provide competitive performance in tasks such as denoising, demixing, and compressive sensing.

\*\*\*\*\*

#### Zero-shot Dual Machine Translation

Lierni Sestorain, Massimiliano Ciaramita, Christian Buck, Thomas Hofmann

Neural Machine Translation (NMT) systems rely on large amounts of parallel data. This is a major challenge for low-resource languages. Building on recent work on unsupervised and semi-supervised methods, we present an approach that combines zero-shot and dual learning. The latter relies on reinforcement learning, to exploit the duality of the machine translation task, and requires only monolingual data

for the target language pair. Experiments on the UN corpus show that a zero-shot dual system, trained on English-French and English-Spanish, outperforms by large margins a standard NMT system in zero-shot translation performance on Spanish-French (both directions). We also evaluate on newstest2014. These experiments show that the zero-shot dual method outperforms the LSTM-based unsupervised NMT system proposed in (Lample et al., 2018b), on the en→fr task, while on the fr→en task it outperforms both the LSTM-based and the Transformers-based unsupervised NMT systems.

\*\*\*\*\*

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

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.

\*\*\*\*\*

MILE: A Multi-Level Framework for Scalable Graph Embedding

Jiongqian Liang, Saket Gurukar, Srinivasan Parthasarathy

Recently there has been a surge of interest in designing graph embedding methods. Few, if any, can scale to a large-sized graph with millions of nodes due to both computational complexity and memory requirements. In this paper, we relax this limitation by introducing the Multi-Level Embedding (MILE) framework - a generic methodology allowing contemporary graph embedding methods to scale to large graphs. MILE repeatedly coarsens the graph into smaller ones using a hybrid matching technique to maintain the backbone structure of the graph. It then applies existing embedding methods on the coarsest graph and refines the embeddings to the original graph through a novel graph convolution neural network that it learns. The proposed MILE framework is agnostic to the underlying graph embedding techniques and can be applied to many existing graph embedding methods without modifying them. We employ our framework on several popular graph embedding techniques and conduct embedding for real-world graphs. Experimental results on five large-scale datasets demonstrate that MILE significantly boosts the speed (order of magnitude) of graph embedding while also often generating embeddings of better quality for the task of node classification. MILE can comfortably scale to a graph with 9 million nodes and 40 million edges, on which existing methods run out of memory or take too long to compute on a modern workstation.

\*\*\*\*\*

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

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 summa



tion 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.

\*\*\*\*\*

Geometry aware convolutional filters for omnidirectional images representation  
Renata Khasanova, Pascal Frossard

Due to their wide field of view, omnidirectional cameras are frequently used by autonomous vehicles, drones and robots for navigation and other computer vision tasks. The images captured by such cameras, are often analysed and classified with techniques designed for planar images that unfortunately fail to properly handle the native geometry of such images. That results in suboptimal performance, and lack of truly meaningful visual features. In this paper we aim at improving popular deep convolutional neural networks so that they can properly take into account the specific properties of omnidirectional data. In particular we propose an algorithm that adapts convolutional layers, which often serve as a core building block of a CNN, to the properties of omnidirectional images. Thus, our filters have a shape and size that adapts with the location on the omnidirectional image. We show that our method is not limited to spherical surfaces and is able to incorporate the knowledge about any kind of omnidirectional geometry inside the deep learning network. As depicted by our experiments, our method outperforms the existing deep neural network techniques for omnidirectional image classification and compression tasks.

\*\*\*\*\*

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

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.

\*\*\*\*\*

Improving Sample-based Evaluation for Generative Adversarial Networks

Shaohui Liu\*, Yi Wei\*, Jiwen Lu, Jie Zhou

In this paper, we propose an improved quantitative evaluation framework for Generative Adversarial Networks (GANs) on generating domain-specific images, where we improve conventional evaluation methods on two levels: the feature representation and the evaluation metric. Unlike most existing evaluation frameworks which transfer the representation of ImageNet inception model to map images onto the feature space, our framework uses a specialized encoder to acquire fine-grained domain-specific representation. Moreover, for datasets with multiple classes, we propose Class-Aware Frechet Distance (CAFD), which employs a Gaussian mixture model on the feature space to better fit the multi-manifold feature distribution. Experiments and analysis on both the feature level and the image level were conducted to demonstrate improvements of our proposed framework over the recently proposed state-of-the-art FID method. To our best knowledge, we are the first to provide counter examples where FID gives inconsistent results with human judgments. It is shown in the experiments that our framework is able to overcome the shortness of FID and improves robustness. Code will be made available.

\*\*\*\*\*

## 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

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

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.

\*\*\*\*\*

## Learn From Neighbour: A Curriculum That Train Low Weighted Samples By Imitating

Benyuan Sun, Yizhou Wang

Deep neural networks, which gain great success in a wide spectrum of applications, are often time, compute and storage hungry. Curriculum learning proposed to boost training of network by a syllabus from easy to hard. However, the relationship between data complexity and network training is unclear: why hard example harm the performance at beginning but helps at end. In this paper, we aim to investigate on this problem. Similar to internal covariate shift in network forward pass, the distribution changes in weight of top layers also affects training of preceding layers during the backward pass. We call this phenomenon inverse "internal covariate shift". Training hard examples aggravates the distribution shifting and damages the training. To address this problem, we introduce a curriculum loss that consists of two parts: a) an adaptive weight that mitigates large early punishment; b) an additional representation loss for low weighted samples. The intuition of the loss is very simple. We train top layers on "good" samples to r

reduce large shifting, and encourage "bad" samples to learn from "good" sample. In detail, the adaptive weight assigns small values to hard examples, reducing the influence of noisy gradients. On the other hand, the less-weighted hard sample receives the proposed representation loss. Low-weighted data gets nearly no training signal and can stuck in embedding space for a long time. The proposed representation loss aims to encourage their training. This is done by letting them learn a better representation from its superior neighbours but not participate in learning of top layers. In this way, the fluctuation of top layers is reduced and hard samples also received signals for training. We found in this paper that curriculum learning needs random sampling between tasks for better training. Our curriculum loss is easy to combine with existing stochastic algorithms like SGD. Experimental result shows an consistent improvement over several benchmark datasets.

\*\*\*\*\*

#### Characterizing Malicious Edges targeting on Graph Neural Networks

Xiaojun Xu, Yue Yu, Bo Li, Le Song, Chengfeng Liu, Carl Gunter

Deep neural networks on graph structured data have shown increasing success in various applications. However, due to recent studies about vulnerabilities of machine learning models, researchers are encouraged to explore the robustness of graph neural networks (GNNs). So far there are two work targeting to attack GNNs by adding/deleting edges to fool graph based classification tasks. Such attacks are challenging to be detected since the manipulation is very subtle compared with traditional graph attacks. In this paper we propose the first detection mechanism against these two proposed attacks. Given a perturbed graph, we propose a novel graph generation method together with link prediction as preprocessing to detect potential malicious edges. We also propose novel features which can be leveraged to perform outlier detection when the number of added malicious edges are large. Different detection components are proposed and tested, and we also evaluate the performance of final detection pipeline. Extensive experiments are conducted to show that the proposed detection mechanism can achieve AUC above 90% against the two attack strategies on both Cora and Citeseer datasets. We also provide in-depth analysis of different attack strategies and corresponding suitable detection methods. Our results shed light on several principles for detecting different types of attacks.

\*\*\*\*\*

#### Convergence Guarantees for RMSProp and ADAM in Non-Convex Optimization and an Empirical Comparison to Nesterov Acceleration

Soham De, Anirbit Mukherjee, Enayat Ullah

RMSProp and ADAM continue to be extremely popular algorithms for training neural nets but their theoretical convergence properties have remained unclear. Further, recent work has seemed to suggest that these algorithms have worse generalization properties when compared to carefully tuned stochastic gradient descent or its momentum variants. In this work, we make progress towards a deeper understanding of ADAM and RMSProp in two ways. First, we provide proofs that these adaptive gradient algorithms are guaranteed to reach criticality for smooth non-convex objectives, and we give bounds on the running time.

Next we design experiments to empirically study the convergence and generalization properties of RMSProp and ADAM against Nesterov's Accelerated Gradient method on a variety of common autoencoder setups and on VGG-9 with CIFAR-10. Through these experiments we demonstrate the interesting sensitivity that ADAM has to its momentum parameter  $\beta_1$ . We show that at very high values of the momentum parameter ( $\beta_1 = 0.99$ ) ADAM outperforms a carefully tuned NAG on most of our experiments, in terms of getting lower training and test losses. On the other hand, NAG can sometimes do better when ADAM's  $\beta_1$  is set to the most commonly used value:  $\beta_1 = 0.9$ , indicating the importance of tuning the hyperparameters of ADAM to get better generalization performance.

We also report experiments on different autoencoders to demonstrate that NAG has better abilities in terms of reducing the gradient norms, and it also produces

iterates which exhibit an increasing trend for the minimum eigenvalue of the Hessian of the loss function at the iterates.

\*\*\*\*\*

#### Graph U-Net

Hongyang Gao, Shuiwang Ji

We consider the problem of representation learning for graph data. Convolutional neural networks can naturally operate on images, but have significant challenges in dealing with graph data. Given images are special cases of graphs with nodes lie on 2D lattices, graph embedding tasks have a natural correspondence with image pixel-wise prediction tasks such as segmentation. While encoder-decoder architectures like U-Net have been successfully applied on many image pixel-wise prediction tasks, similar methods are lacking for graph data. This is due to the fact that pooling and up-sampling operations are not natural on graph data. To address these challenges, we propose novel graph pooling (gPool) and unpooling (gUnpool) operations in this work. The gPool layer adaptively selects some nodes to form a smaller graph based on their scalar projection values on a trainable projection vector. We further propose the gUnpool layer as the inverse operation of the gPool layer. The gUnpool layer restores the graph into its original structure using the position information of nodes selected in the corresponding gPool layer. Based on our proposed gPool and gUnpool layers, we develop an encoder-decoder model on graph, known as the graph U-Net. Our experimental results on node classification tasks demonstrate that our methods achieve consistently better performance than previous models.

\*\*\*\*\*

#### Learning Representations in Model-Free Hierarchical Reinforcement Learning

Jacob Rafati, David Noelle

Common approaches to Reinforcement Learning (RL) are seriously challenged by large-scale applications involving huge state spaces and sparse delayed reward feedback. Hierarchical Reinforcement Learning (HRL) methods attempt to address this scalability issue by learning action selection policies at multiple levels of temporal abstraction. Abstraction can be had by identifying a relatively small set of states that are likely to be useful as subgoals, in concert with the learning of corresponding skill policies to achieve those subgoals. Many approaches to subgoal discovery in HRL depend on the analysis of a model of the environment, but the need to learn such a model introduces its own problems of scale. Once subgoals are identified, skills may be learned through intrinsic motivation, introducing an internal reward signal marking subgoal attainment. In this paper, we present a novel model-free method for subgoal discovery using incremental unsupervised learning over a small memory of the most recent experiences of the agent. When combined with an intrinsic motivation learning mechanism, this method learns subgoals and skills together, based on experiences in the environment. Thus, we offer an original approach to HRL that does not require the acquisition of a model of the environment, suitable for large-scale applications. We demonstrate the efficiency of our method on two RL problems with sparse delayed feedback: a variant of the rooms environment and the ATARI 2600 game called Montezuma's Revenge.

\*\*\*\*\*

#### When Will Gradient Methods Converge to Max-margin Classifier under ReLU Models?

Tengyu Xu, Yi Zhou, Kaiyi Ji, Yingbin Liang

We study the implicit bias of gradient descent methods in solving a binary classification problem over a linearly separable dataset. The classifier is described by a nonlinear ReLU model and the objective function adopts the exponential loss function. We first characterize the landscape of the loss function and show that there can exist spurious asymptotic local minima besides asymptotic global minima. We then show that gradient descent (GD) can converge to either a global or a local max-margin direction, or may diverge from the desired max-margin direction in a general context. For stochastic gradient descent (SGD), we show that it converges in expectation to either the global or the local max-margin direction if SGD converges. We further explore the implicit bias of these algorithms in 1

earning a multi-neuron network under certain stationary conditions, and show that the learned classifier maximizes the margins of each sample pattern partition under the ReLU activation.

\*\*\*\*\*

Feature Transformers: A Unified Representation Learning Framework for Lifelong Learning

Hariharan Ravishankar, Rahul Venkataramani, Saihareesh Anamandra, Prasad Sudhakar

Despite the recent advances in representation learning, lifelong learning continues

to be one of the most challenging and unconquered problems. Catastrophic forgetting

and data privacy constitute two of the important challenges for a successful lifelong learner. Further, existing techniques are designed to handle only specific

manifestations of lifelong learning, whereas a practical lifelong learner is expected

to switch and adapt seamlessly to different scenarios. In this paper, we present a

single, unified mathematical framework for handling the myriad variants of lifelong

learning, while alleviating these two challenges. We utilize an external memory to store only the features representing past data and learn richer and newer representations incrementally through transformation neural networks - feature transformers. We define, simulate and demonstrate exemplary performance on a realistic lifelong experimental setting using the MNIST rotations dataset, paving

the way for practical lifelong learners. To illustrate the applicability of our method

in data sensitive domains like healthcare, we study the pneumothorax classification

problem from X-ray images, achieving near gold standard performance.

We also benchmark our approach with a number of state-of-the-art methods on MNIST rotations and iCIFAR100 datasets demonstrating superior performance.

\*\*\*\*\*

Stability of Stochastic Gradient Method with Momentum for Strongly Convex Loss Functions

Ali Ramezani-Kebrya, Ashish Khisti, and Ben Liang

While momentum-based methods, in conjunction with the stochastic gradient descent, are widely used when training machine learning models, there is little theoretical understanding on the generalization error of such methods. In practice, the momentum parameter is often chosen in a heuristic fashion with little theoretical guidance. In this work, we use the framework of algorithmic stability to provide an upper-bound on the generalization error for the class of strongly convex loss functions, under mild technical assumptions. Our bound decays to zero inversely with the size of the training set, and increases as the momentum parameter is increased. We also develop an upper-bound on the expected true risk, in terms of the number of training steps, the size of the training set, and the momentum parameter.

\*\*\*\*\*

Learned optimizers that outperform on wall-clock and validation loss

Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, Jascha Sohl-dickstein

Deep learning has shown that learned functions can dramatically outperform hand-designed functions on perceptual tasks. Analogously, this suggests that learned update functions may similarly outperform current hand-designed optimizers, especially for specific tasks. However, learned optimizers are notoriously difficult to train and have yet to demonstrate wall-clock speedups over hand-designed optimizers, and thus are rarely used in practice. Typically, learned optimizers are trained by truncated backpropagation through an unrolled optimization process.

The resulting gradients are either strongly biased (for short truncations) or have exploding norm (for long truncations). In this work we propose a training sch

eme which overcomes both of these difficulties, by dynamically weighting two unbiased gradient estimators for a variational loss on optimizer performance. This allows us to train neural networks to perform optimization faster than well tuned first-order methods. Moreover, by training the optimizer against validation loss, as opposed to training loss, we are able to use it to train models which generalize better than those trained by first order methods. We demonstrate these results on problems where our learned optimizer trains convolutional networks in a fifth of the wall-clock time compared to tuned first-order methods, and with a n improvement

\*\*\*\*\*

An Empirical study of Binary Neural Networks' Optimisation

Milad Alizadeh, Javier Fernández-Marqués, Nicholas D. Lane, Yarin Gal

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.

\*\*\*\*\*

Generative Ensembles for Robust Anomaly Detection

Hyunsun Choi, Eric Jang

Deep generative models are capable of learning probability distributions over large, high-dimensional datasets such as images, video and natural language. Generative models trained on samples from  $p(x)$  ought to assign low likelihoods to out-of-distribution (OOD) samples from  $q(x)$ , making them suitable for anomaly detection applications. We show that in practice, likelihood models are themselves susceptible to OOD errors, and even assign large likelihoods to images from other natural datasets. To mitigate these issues, we propose Generative Ensembles, a model-independent technique for OOD detection that combines density-based anomaly detection with uncertainty estimation. Our method outperforms ODIN and VIB baselines on image datasets, and achieves comparable performance to a classification model on the Kaggle Credit Fraud dataset.

\*\*\*\*\*

Unsupervised Learning of Sentence Representations Using Sequence Consistency

Siddhartha Brahma

Computing universal distributed representations of sentences is a fundamental task in natural language processing. We propose ConsSent, a simple yet surprisingly powerful unsupervised method to learn such representations by enforcing consistency constraints on sequences of tokens. We consider two classes of such constraints - sequences that form a sentence and between two sequences that form a sentence when merged. We learn sentence encoders by training them to distinguish between consistent and inconsistent examples, the latter being generated by randomly perturbing consistent examples in six different ways. Extensive evaluation on several transfer learning and linguistic probing tasks shows improved performance over strong unsupervised and supervised baselines, substantially surpassing them in several cases. Our best results are achieved by training sentence encoders in a multitask setting and by an ensemble of encoders trained on the individual tasks.

\*\*\*\*\*

DADAM: A consensus-based distributed adaptive gradient method for online optimization

Parvin Nazari, Davoud Ataee Tarzanagh, George Michailidis

Online and stochastic optimization methods such as SGD, ADAGRAD and ADAM are key algorithms in solving large-scale machine learning problems including deep learning. A number of schemes that are based on communications of nodes with a central server have been recently proposed in the literature to parallelize them. A bottleneck of such centralized algorithms lies on the high communication cost incurred by the central node. In this paper, we present a new consensus-based distributed adaptive moment estimation method (DADAM) for online optimization over a decentralized network that enables data parallelization, as well as decentralized computation. Such a framework not only can be extremely useful for learning agents with access to only local data in a communication constrained environment, but as shown in this work also outperform centralized adaptive algorithms such as ADAM for certain realistic classes of loss functions. We analyze the convergence properties of the proposed algorithm and provide a  $\text{dynamic regret}$  bound on the convergence rate of adaptive moment estimation methods in both stochastic and deterministic settings. Empirical results demonstrate that DADAM works well in practice and compares favorably to competing online optimization methods.

\*\*\*\*\*

A2BCD: Asynchronous Acceleration with Optimal Complexity

Robert Hannah, Fei Feng, Wotao Yin

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.

\*\*\*\*\*

Bayesian Deep Learning via Stochastic Gradient MCMC with a Stochastic Approximation Adaptation

Wei Deng, Xiao Zhang, Faming Liang, Guang Lin

We propose a robust Bayesian deep learning algorithm to infer complex posteriors with latent variables. Inspired by dropout, a popular tool for regularization and model ensemble, we assign sparse priors to the weights in deep neural networks (DNN) in order to achieve automatic "dropout" and avoid over-fitting. By alternatively sampling from posterior distribution through stochastic gradient Markov Chain Monte Carlo (SG-MCMC) and optimizing latent variables via stochastic approximation (SA), the trajectory of the target weights is proved to converge to the true posterior distribution conditioned on optimal latent variables. This ensures a stronger regularization on the over-fitted parameter space and more accurate uncertainty quantification on the decisive variables. Simulations from large-p-small-n regressions showcase the robustness of this method when applied to models with latent variables. Additionally, its application on the convolutional neural networks (CNN) leads to state-of-the-art performance on MNIST and Fashion MNIST datasets and improved resistance to adversarial attacks.

\*\*\*\*\*

Detecting Memorization in ReLU Networks

Edo Collins, Siavash Arjomand Bigdeli, Sabine Ssstrunk

We propose a new notion of 'non-linearity' of a network layer with respect to an input batch that is based on its proximity to a linear system, which is reflected in the non-negative rank of the activation matrix.

We measure this non-linearity by applying non-negative factorization to the acti

vation matrix.

Considering batches of similar samples, we find that high non-linearity in deep layers is indicative of memorization. Furthermore, by applying our approach layer-by-layer, we find that the mechanism for memorization consists of distinct phases. We perform experiments on fully-connected and convolutional neural networks trained on several image and audio datasets. Our results demonstrate that as an indicator for memorization, our technique can be used to perform early stopping

.

\*\*\*\*\*

LSH Microbatches for Stochastic Gradients: Value in Rearrangement

Eliav Buchnik, Edith Cohen, Avinatan Hassidim, Yossi Matias

Metric embeddings are immensely useful representations of associations between entities (images, users, search queries, words, and more). Embeddings are learned by optimizing a loss objective of the general form of a sum over example associations. Typically, the optimization uses stochastic gradient updates over minibatches of examples that are arranged independently at random. In this work, we propose the use of {\em structured arrangements} through randomized {\em microbatches} of examples that are more likely to include similar ones. We make a principled argument for the properties of our arrangements that accelerate the training and present efficient algorithms to generate microbatches that respect the marginal distribution of training examples. Finally, we observe experimentally that our structured arrangements accelerate training by 3-20\%. Structured arrangements emerge as a powerful and novel performance knob for SGD that is independent and complementary to other SGD hyperparameters and thus is a candidate for wide deployment.

\*\*\*\*\*

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

Thomas Miconi, Aditya Rawal, Jeff Clune, Kenneth O. Stanley

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.

\*\*\*\*\*

Hierarchical Bayesian Modeling for Clustering Sparse Sequences in the Context of Group Profiling

Ishani Chakraborty

This paper proposes a hierarchical Bayesian model for clustering sparse sequences. This is a mixture model and does not need the data to be represented by a Gaussian mixture and that gives significant modelling freedom. It also generates a very interpretable profile for the discovered latent groups. The data that was used for the work have been contributed by a restaurant loyalty program company. The data is a collection of sparse sequences where each entry of each sequence is the number of user visits of one week to some restaurant. This algorithm successfully clustered the data and calculated the expected user affiliation in each cluster.

\*\*\*\*\*

Discovery of Natural Language Concepts in Individual Units of CNNs

Seil Na, Yo Joong Choe, Dong-Hyun Lee, Gunhee Kim



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.

\*\*\*\*\*

#### Neural Distribution Learning for generalized time-to-event prediction

Egil Martinsson, Adrian Kim, Jaesung Huh, Jaegul Choo, Jung-Woo Ha

Predicting the time to the next event is an important task in various domains. However, due to censoring and irregularly sampled sequences, time-to-event prediction has resulted in limited success only for particular tasks, architectures and data. Using recent advances in probabilistic programming and density networks, we make the case for a generalized parametric survival approach, sequentially predicting a distribution over the time to the next event. Unlike previous work, the proposed method can use asynchronously sampled features for censored, discrete, and multivariate data. Furthermore, it achieves good performance and near perfect calibration for probabilistic predictions without using rigid network-architectures, multitask approaches, complex learning schemes or non-trivial adaptations of cox-models. We firmly establish that this can be achieved in the standard neural network framework by simply switching out the output layer and loss function.

\*\*\*\*\*

#### Learning Multi-Level Hierarchies with Hindsight

Andrew Levy, George Konidaris, Robert Platt, Kate Saenko

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.

\*\*\*\*\*

#### Learning Hash Codes via Hamming Distance Targets

Martin Loncaric, Ryan Weber, Bowei Liu

We present a powerful new loss function and training scheme for learning binary hash codes with any differentiable model and similarity function. Our loss function improves over prior methods by using log likelihood loss on top of an accurate approximation for the probability that two inputs fall within a Hamming distance target. Our novel training scheme obtains a good estimate of the true gradient by better sampling inputs and evaluating loss terms between all pairs of inputs in each minibatch.

inibatch.

To fully leverage the resulting hashes, we use multi-indexing.

We demonstrate that these techniques provide large improvements to a similarity search tasks.

We report the best results to date on competitive information retrieval tasks for Imagenet and SIFT 1M, improving recall from 73% to 85% and reducing query cost by a factor of 2-8, respectively.

\*\*\*\*\*

Fortified Networks: Improving the Robustness of Deep Networks by Modeling the Manifold of Hidden Representations

Alex Lamb, Jonathan Binas, Anirudh Goyal, Dmitriy Serdyuk, Sandeep Subramanian, Ioannis Mitliagkas, Yoshua Bengio

Deep networks have achieved impressive results across a variety of important tasks. However, a known weakness is a failure to perform well when evaluated on data which differ from the training distribution, even if these differences are very small, as is the case with adversarial examples. We propose *Fortified Networks*, a simple transformation of existing networks, which "fortifies" the hidden layers in a deep network by identifying when the hidden states are off of the data manifold, and maps these hidden states back to parts of the data manifold where the network performs well. Our principal contribution is to show that fortifying these hidden states improves the robustness of deep networks and our experiments (i) demonstrate improved robustness to standard adversarial attacks in both black-box and white-box threat models; (ii) suggest that our improvements are not primarily due to the problem of deceptively good results due to degraded quality in the gradient signal (the gradient masking problem) and (iii) show the advantage of doing this fortification in the hidden layers instead of the input space. We demonstrate improvements in adversarial robustness on three datasets (MNIST, Fashion MNIST, CIFAR10), across several attack parameters, both white-box and black-box settings, and the most widely studied attacks (FGSM, PGD, Carlini-Wagner). We show that these improvements are achieved across a wide variety of hyperparameters.

\*\*\*\*\*

ProMP: Proximal Meta-Policy Search

Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, Pieter Abbeel

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.

\*\*\*\*\*

CDeepEx: Contrastive Deep Explanations

Amir Feghahati, Christian R. Shelton, Michael J. Pazzani, Kevin Tang

We propose a method which can visually explain the classification decision of deep neural networks (DNNs). There are many proposed methods in machine learning and computer vision seeking to clarify the decision of machine learning black boxes, specifically DNNs. All of these methods try to gain insight into why the network "chose class A" as an answer. Humans, when searching for explanations, ask two types of questions. The first question is, "Why did you choose this answer?" The second question asks, "Why did you not choose answer B over A?" The previously proposed methods are either not able to provide the latter directly or efficiently.

We introduce a method capable of answering the second question both directly and efficiently. In this work, we limit the inputs to be images. In general, the proposed method generates explanations in the input space of any model capable of efficient evaluation and gradient evaluation. We provide results, showing the superiority of this approach for gaining insight into the inner representation of machine learning models.

\*\*\*\*\*

#### 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

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 or 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.

\*\*\*\*\*

#### Mixture of Pre-processing Experts Model for Noise Robust Deep Learning on Resource Constrained Platforms

Taesik Na, Minah Lee, Burhan A. Mudassar, Priyabrata Saha, Jong Hwan Ko, Saibal Mukhopadhyay

Deep learning on an edge device requires energy efficient operation due to ever diminishing power budget. Intentional low quality data during the data acquisition for longer battery life, and natural noise from the low cost sensor degrade the quality of target output which hinders adoption of deep learning on an edge device. To overcome these problems, we propose simple yet efficient mixture of pre-processing experts (MoPE) model to handle various image distortions including low resolution and noisy images. We also propose to use adversarially trained auto encoder as a pre-processing expert for the noisy images. We evaluate our proposed method for various machine learning tasks including object detection on MS-COCO 2014 dataset, multiple object tracking problem on MOT-Challenge dataset, and human activity recognition on UCF 101 dataset. Experimental results show that the proposed method achieves better detection, tracking and activity recognition accuracies under noise without sacrificing accuracies for the clean images. The overheads of our proposed MoPE are 0.67% and 0.17% in terms of memory and computation compared to the baseline object detection network.

\*\*\*\*\*

#### 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

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 extensive suite of 19 levels of increasing difficulty. Each level gradually leads the agent towards acquiring a combinatorially rich synthetic language, which is a pr

oper 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.

\*\*\*\*\*

#### PAIRWISE AUGMENTED GANS WITH ADVERSARIAL RECONSTRUCTION LOSS

Aibek Alanov, Max Kochurov, Daniil Yashkov, Dmitry Vetrov

We propose a novel autoencoding model called Pairwise Augmented GANs. We train a generator and an encoder jointly and in an adversarial manner. The generator network learns to sample realistic objects. In turn, the encoder network at the same time is trained to map the true data distribution to the prior in latent space. To ensure good reconstructions, we introduce an augmented adversarial reconstruction loss. Here we train a discriminator to distinguish two types of pairs: an object with its augmentation and the one with its reconstruction. We show that such adversarial loss compares objects based on the content rather than on the exact match. We experimentally demonstrate that our model generates samples and reconstructions of quality competitive with state-of-the-art on datasets MNIST, CIFAR10, CelebA and achieves good quantitative results on CIFAR10.

\*\*\*\*\*

#### Guaranteed Recovery of One-Hidden-Layer Neural Networks via Cross Entropy

Haoyu Fu, Yuejie Chi, Yingbin Liang

We study model recovery for data classification, where the training labels are generated from a one-hidden-layer fully-connected neural network with sigmoid activations, and the goal is to recover the weight vectors of the neural network. We prove that under Gaussian inputs, the empirical risk function using cross entropy exhibits strong convexity and smoothness uniformly in a local neighborhood of the ground truth, as soon as the sample complexity is sufficiently large. This implies that if initialized in this neighborhood, which can be achieved via the tensor method, gradient descent converges linearly to a critical point that is provably close to the ground truth without requiring a fresh set of samples at each iteration. To the best of our knowledge, this is the first global convergence guarantee established for the empirical risk minimization using cross entropy via gradient descent for learning one-hidden-layer neural networks, at the near-optimal sample and computational complexity with respect to the network input dimension.

\*\*\*\*\*

#### Slimmable Neural Networks

Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, Thomas Huang

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](https://github.com/JiahuiYu/slimmable_networks)

\*\*\*\*\*

#### Entropic GANs meet VAEs: A Statistical Approach to Compute Sample Likelihoods in GANs

Yogesh Balaji, Hamed Hasani, Rama Chellappa, Soheil Feizi

Building on the success of deep learning, two modern approaches to learn a proba

bility model of the observed data are Generative Adversarial Networks (GANs) and Variational AutoEncoders (VAEs). VAEs consider an explicit probability model for the data and compute a generative distribution by maximizing a variational lower-bound on the log-likelihood function. GANs, however, compute a generative model by minimizing a distance between observed and generated probability distributions without considering an explicit model for the observed data. The lack of having explicit probability models in GANs prohibits computation of sample likelihoods in their frameworks and limits their use in statistical inference problems.

In this work, we show that an optimal transport GAN with the entropy regularization can be viewed as a generative model that maximizes a lower-bound on average sample likelihoods, an approach that VAEs are based on. In particular, our proof constructs an explicit probability model for GANs that can be used to compute likelihood statistics within GAN's framework. Our numerical results on several datasets demonstrate consistent trends with the proposed theory.

\*\*\*\*\*

#### Learning Self-Imitating Diverse Policies

Tanmay Gangwani, Qiang Liu, Jian Peng

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 difficulty 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

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 $\times$  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.

\*\*\*\*\*

#### Visual Imitation Learning with Recurrent Siamese Networks

Glen Berseth, Christopher J. Pal

People are incredibly skilled at imitating others by simply observing them. They achieve this even in the presence of significant morphological differences and capabilities. Further, people are able to do this from raw perceptions of the actions of others, without direct access to the abstracted demonstration actions and with only partial state information. People therefore solve a difficult problem of understanding the salient features of both observations of others and the relationship to their own state when learning to imitate specific tasks.

However, we can attempt to reproduce a similar demonstration via trial and error and through this gain more understanding of the task space.

To reproduce this ability an agent would need to both learn how to recognize the differences between itself and some demonstration and at the same time learn to minimize the distance between its own performance and that of the demonstration.

In this paper we propose an approach using only visual information to learn a distance metric between agent behaviour and a given video demonstration.

We train an RNN-based siamese model to compute distances in space and time between motion clips while training an RL policy to minimize this distance.

Furthermore, we examine a particularly challenging form of this problem where the agent must learn an imitation based task given a single demonstration.

We demonstrate our approach in the setting of deep learning based control for physical simulation of humanoid walking in both 2D with 10 degrees of freedom (DoF) and 3D with 38 DoF.

\*\*\*\*\*

#### A NON-LINEAR THEORY FOR SENTENCE EMBEDDING

Hichem Mezaoui, Isar Nejadgholi

This paper revisits the Random Walk model for sentence embedding in the context of non-extensive statistics. We propose a non-extensive algebra to compute the discourse vector. We argue that by doing so we are taking into account high non-linearity in the semantic space. Furthermore, we show that by considering a non-extensive algebra, the compounding effect of the vector length is mitigated. Overall, we show that the proposed model leads to good sentence embedding. We evaluate the embedding method on textual similarity tasks.

\*\*\*\*\*

#### Sufficient Conditions for Robustness to Adversarial Examples: a Theoretical and Empirical Study with Bayesian Neural Networks

Yarin Gal, Lewis Smith

We prove, under two sufficient conditions, that idealised models can have no adversarial examples. We discuss which idealised models satisfy our conditions, and show that idealised Bayesian neural networks (BNNs) satisfy these. We continue by studying near-idealised BNNs using HMC inference, demonstrating the theoretical ideas in practice. We experiment with HMC on synthetic data derived from MNIST for which we know the ground-truth image density, showing that near-perfect epistemic uncertainty correlates to density under image manifold, and that adversarial images lie off the manifold in our setting. This suggests why MC dropout, which can be seen as performing approximate inference, has been observed to be an effective defence against adversarial examples in practice; We highlight failure-cases of non-idealised BNNs relying on dropout, suggesting a new attack for dropout models and a new defence as well. Lastly, we demonstrate the defence on a cats-vs-dogs image classification task with a VGG13 variant.

\*\*\*\*\*

#### Dual Learning: Theoretical Study and Algorithmic Extensions

Zhibing Zhao, Yingce Xia, Tao Qin, Tie-Yan Liu

Dual learning has been successfully applied in many machine learning applications, including machine translation, image-to-image transformation, etc. The high-level idea of dual learning is very intuitive: if we map an  $x$  from one domain to another and then map it back, we should recover the original  $x$ . Although its effectiveness has been empirically verified, theoretical understanding of dual learning is still missing. In this paper, we conduct a theoretical study to understand why and when dual learning can improve a mapping function. Based on the theoretical discoveries, we extend dual learning by introducing more related mapping

s and propose highly symmetric frameworks, cycle dual learning and multipath dual learning, in both of which we can leverage the feedback signals from additional domains to improve the qualities of the mappings. We prove that both cycle dual learning and multipath dual learning can boost the performance of standard dual learning under mild conditions. Experiments on WMT 14 English $\leftrightarrow$ German and Multi UN English $\leftrightarrow$ French translations verify our theoretical findings on dual learning, and the results on the translations among English, French, and Spanish of Multi UN demonstrate the efficacy of cycle dual learning and multipath dual learning.

\*\*\*\*\*

UaiNets: From Unsupervised to Active Deep Anomaly Detection

Tiago Pimentel, Marianne Monteiro, Juliano Viana, Adriano Veloso, Nivio Ziviani

This work presents a method for active anomaly detection which can be built upon existing deep learning solutions for unsupervised anomaly detection. We show that a prior needs to be assumed on what the anomalies are, in order to have performance guarantees in unsupervised anomaly detection. We argue that active anomaly detection has, in practice, the same cost of unsupervised anomaly detection but with the possibility of much better results. To solve this problem, we present a new layer that can be attached to any deep learning model designed for unsupervised anomaly detection to transform it into an active method, presenting results on both synthetic and real anomaly detection datasets.

\*\*\*\*\*

Deep Layers as Stochastic Solvers

Adel Bibi, Bernard Ghanem, Vladlen Koltun, Rene Ranftl

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

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

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.

\*\*\*\*\*

## DEEP HIERARCHICAL MODEL FOR HIERARCHICAL SELECTIVE CLASSIFICATION AND ZERO SHOT LEARNING

Eliyahu Sason, Koby Crammer

Object recognition in real-world image scenes is still an open problem. With the growing number of classes, the similarity structures between them become complex and the distinction between classes blurs, which makes the classification problem particularly challenging. Standard N-way discrete classifiers treat all classes as disconnected and unrelated, and therefore unable to learn from their semantic relationships. In this work, we present a hierarchical inter-class relationship model and train it using a newly proposed probability-based loss function. Our hierarchical model provides significantly better semantic generalization ability compared to a regular N-way classifier. We further proposed an algorithm where given a probabilistic classification model it can return the input corresponding super-group based on classes hierarchy without any further learning. We deploy it in two scenarios in which super-group retrieval can be useful. The first one, selective classification, deals with the problem of low-confidence classification, wherein a model is unable to make a successful exact classification. The second, zero-shot learning problem deals with making reasonable inferences on novel classes. Extensive experiments with the two scenarios show that our proposed hierarchical model yields more accurate and meaningful super-class predictions compared to a regular N-way classifier because of its significantly better semantic generalization ability.

\*\*\*\*\*

A bird's eye view on coherence, and a worm's eye view on cohesion

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujuan Li, Mengdi Wang, Jianfeng Gao

Generating coherent and cohesive long-form texts is a challenging problem in natural language generation. Previous works relied on a large amount of human-generated texts to train neural language models, however, few attempted to explicitly model the desired linguistic properties of natural language text, such as coherence and cohesion using neural networks. In this work, we train two expert discriminators for coherence and cohesion to provide hierarchical feedback for text generation. We also propose a simple variant of policy gradient, called 'negative-critical sequence training' in which the reward 'baseline' is constructed from randomly generated negative samples. We demonstrate the effectiveness of our approach through empirical studies, showing improvements over the strong baseline -- attention-based bidirectional MLE-trained neural language model -- in a number of automated metrics. The proposed model can serve as baseline architectures to promote further research in modeling additional linguistic properties for downstream NLP tasks.

\*\*\*\*\*

Unsupervised Discovery of Parts, Structure, and Dynamics

Zhenjia Xu\*, Zhijian Liu\*, Chen Sun, Kevin Murphy, William T. Freeman, Joshua B. Tenenbaum, Jiajun Wu

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.

\*\*\*\*\*

#### State-Denoised Recurrent Neural Networks

Michael C. Mozer, Denis Kazakov, Robert V. Lindsey

Recurrent neural networks (RNNs) are difficult to train on sequence processing tasks, not only because input noise may be amplified through feedback, but also because any inaccuracy in the weights has similar consequences as input noise. We describe a method for denoising the hidden state during training to achieve more robust representations thereby improving generalization performance. Attractor dynamics are incorporated into the hidden state to 'clean up' representations at each step of a sequence. The attractor dynamics are trained through an auxiliary denoising loss to recover previously experienced hidden states from noisy versions of those states. This state-denoised recurrent neural network (SDRNN) performs multiple steps of internal processing for each external sequence step. On a range of tasks, we show that the SDRNN outperforms a generic RNN as well as a variant of the SDRNN with attractor dynamics on the hidden state but without the auxiliary loss. We argue that attractor dynamics---and corresponding connectivity constraints---are an essential component of the deep learning arsenal and should be invoked not only for recurrent networks but also for improving deep feedforward nets and intertask transfer.

\*\*\*\*\*

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

Wengong Jin, Kevin Yang, Regina Barzilay, Tommi Jaakkola

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

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.

\*\*\*\*\*

#### COCO-GAN: Conditional Coordinate Generative Adversarial Network

Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, Hwann-Tzong Chen

Recent advancements on Generative Adversarial Network (GAN) have inspired a wide range of works that generate synthetic images. However, the current processes have to generate an entire image at once, and therefore resolutions are limited by memory or computational constraints. In this work, we propose COnditional COor

dinate GAN (COCO-GAN), which generates a specific patch of an image conditioned on a spatial position rather than the entire image at a time. The generated patches are later combined together to form a globally coherent full-image. With this process, we show that the generated image can achieve competitive quality to state-of-the-arts and the generated patches are locally smooth between consecutive neighbors. One direct implication of the COCO-GAN is that it can be applied on to any coordinate systems including the cylindrical systems which makes it feasible for generating panorama images. The fact that the patch generation process is independent to each other inspires a wide range of new applications: firstly, "Patch-Inspired Image Generation" enables us to generate the entire image based on a single patch. Secondly, "Partial-Scene Generation" allows us to generate images within a customized target region. Finally, thanks to COCO-GAN's patch generation and massive parallelism, which enables combining patches for generating a full-image with higher resolution than state-of-the-arts.

\*\*\*\*\*

#### Differentially Private Federated Learning: A Client Level Perspective

Robin C. Geyer, Tassilo J. Klein, Moin Nabi

Federated learning is a recent advance in privacy protection.

In this context, a trusted curator aggregates parameters optimized in decentralized fashion by multiple clients. The resulting model is then distributed back to all clients, ultimately converging to a joint representative model without explicitly having to share the data.

However, the protocol is vulnerable to differential attacks, which could originate from any party contributing during federated optimization. In such an attack, a client's contribution during training and information about their data set is revealed through analyzing the distributed model.

We tackle this problem and propose an algorithm for client sided differential privacy preserving federated optimization. The aim is to hide clients' contributions during training, balancing the trade-off between privacy loss and model performance.

Empirical studies suggest that given a sufficiently large number of participating clients, our proposed procedure can maintain client-level differential privacy at only a minor cost in model performance.

\*\*\*\*\*

#### Implicit Autoencoders

Alireza Makhzani

In this paper, we describe the "implicit autoencoder" (IAE), a generative autoencoder in which both the generative path and the recognition path are parametrized by implicit distributions. We use two generative adversarial networks to define the reconstruction and the regularization cost functions of the implicit autoencoder, and derive the learning rules based on maximum-likelihood learning. Using implicit distributions allows us to learn more expressive posterior and conditional likelihood distributions for the autoencoder. Learning an expressive conditional likelihood distribution enables the latent code to only capture the abstract and high-level information of the data, while the remaining information is captured by the implicit conditional likelihood distribution. For example, we show that implicit autoencoders can disentangle the global and local information, and perform deterministic or stochastic reconstructions of the images. We further show that implicit autoencoders can disentangle discrete underlying factors of variation from the continuous factors in an unsupervised fashion, and perform clustering and semi-supervised learning.

\*\*\*\*\*

#### Adversarial Information Factorization

Antonia Creswell, Yumnah Mohamied, Biswa Sengupta, Anil Bharath

We propose a novel generative model architecture designed to learn representations for images that factor out a single attribute from the rest of the representation. A single object may have many attributes which when altered do not change the identity of the object itself. Consider the human face; the identity of a particular person is independent of whether or not they happen to be wearing glasses. The attribute of wearing glasses can be changed without changing the identity

y of the person. However, the ability to manipulate and alter image attributes without altering the object identity is not a trivial task. Here, we are interested in learning a representation of the image that separates the identity of an object (such as a human face) from an attribute (such as 'wearing glasses'). We demonstrate the success of our factorization approach by using the learned representation to synthesize the same face with and without a chosen attribute. We refer to this specific synthesis process as image attribute manipulation. We further demonstrate that our model achieves competitive scores, with state of the art, on a facial attribute classification task.

\*\*\*\*\*

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

Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, Kyogu Lee

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.

\*\*\*\*\*

#### Evaluation Methodology for Attacks Against Confidence Thresholding Models

Ian Goodfellow, Yao Qin, David Berthelot

Current machine learning algorithms can be easily fooled by adversarial examples. One possible solution path is to make models that use confidence thresholding to avoid making mistakes. Such models refuse to make a prediction when they are not confident of their answer. We propose to evaluate such models in terms of tradeoff curves with the goal of high success rate on clean examples and low failure rate on adversarial examples. Existing untargeted attacks developed for models that do not use confidence thresholding tend to underestimate such models' vulnerability. We propose the MaxConfidence family of attacks, which are optimal in a variety of theoretical settings, including one realistic setting: attacks against linear models. Experiments show the attack attains good results in practice. We show that simple defenses are able to perform well on MNIST but not on CIFAR, contributing further to previous calls that MNIST should be retired as a benchmarking dataset for adversarial robustness research. We release code for these evaluations as part of the cleverhans (Papernot et al 2018) library (ICLR reviewers should be careful not to look at who contributed these features to cleverhans to avoid de-anonymizing this submission).

\*\*\*\*\*

#### Infinitely Deep Infinite-Width Networks

Jovana Mitrovic, Peter Wirnsberger, Charles Blundell, Dino Sejdinovic, Yee Whye Teh

Infinite-width neural networks have been extensively used to study the theoretical properties underlying the extraordinary empirical success of standard, finite-width neural networks. Nevertheless, until now, infinite-width networks have been limited to at most two hidden layers. To address this shortcoming, we study the initialisation requirements of these networks and show that the main challenge for constructing them is defining the appropriate sampling distributions for the weights. Based on these observations, we propose a principled approach to weight initialisation that correctly accounts for the functional nature of the hidden layer activations and facilitates the construction of arbitrarily many infinite-width layers, thus enabling the construction of arbitrarily deep infinite-width

th networks. The main idea of our approach is to iteratively reparametrise the hidden-layer activations into appropriately defined reproducing kernel Hilbert spaces and use the canonical way of constructing probability distributions over these spaces for specifying the required weight distributions in a principled way.

Furthermore, we examine the practical implications of this construction for standard, finite-width networks. In particular, we derive a novel weight initialisation scheme for standard, finite-width networks that takes into account the structure of the data and information about the task at hand. We demonstrate the effectiveness of this weight initialisation approach on the MNIST, CIFAR-10 and Year Prediction MSD datasets.

\*\*\*\*\*

## Exploiting Environmental Variation to Improve Policy Robustness in Reinforcement Learning

Siddharth Mysore, Robert Platt, Kate Saenko

Conventional reinforcement learning rarely considers how the physical variations in the environment (eg. mass, drag, etc.) affect the policy learned by the agent. In this paper, we explore how changes in the environment affect policy generalization. We observe experimentally that, for each task we considered, there exists an optimal environment setting that results in the most robust policy that generalizes well to future environments. We propose a novel method to exploit this observation to develop robust actor policies, by automatically developing a sampling curriculum over environment settings to use in training. Ours is a model-free approach and experiments demonstrate that the performance of our method is on par with the best policies found by an exhaustive grid search, while bearing a significantly lower computational cost.

\*\*\*\*\*

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

Frederic Koehler, Andrej Risteski

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  $L_1$  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.

\*\*\*\*\*

## GEOMETRIC AUGMENTATION FOR ROBUST NEURAL NETWORK CLASSIFIERS

Robert M. Taylor, Yusong Tan

We introduce a novel geometric perspective and unsupervised model augmentation framework for transforming traditional deep (convolutional) neural networks into adversarially robust classifiers. Class-conditional probability densities based

on Bayesian nonparametric mixtures of factor analyzers (BNP-MFA) over the input space are used to design soft decision labels for feature to label isometry. Class conditional distributions over features are also learned using BNP-MFA to develop plug-in maximum a posterior (MAP) classifiers to replace the traditional multinomial logistic softmax classification layers. This novel unsupervised augmented framework, which we call geometrically robust networks (GRN), is applied to CIFAR-10, CIFAR-100, and to Radio-ML (a time series dataset for radio modulation recognition). We demonstrate the robustness of GRN models to adversarial attacks from fast gradient sign method, Carlini-Wagner, and projected gradient descent.

\*\*\*\*\*

#### Neural Probabilistic Motor Primitives for Humanoid Control

Josh Merel, Leonard Hasenclever, Alexandre Galashov, Arun Ahuja, Vu Pham, Greg Wayne, Yee Whye Teh, Nicolas Heess

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 to Decompose Compound Questions with Reinforcement Learning

Haihong Yang, Han Wang, Shuang Guo, Wei Zhang, Huajun Chen

As for knowledge-based question answering, a fundamental problem is to relax the assumption of answerable questions from simple questions to compound questions. Traditional approaches firstly detect topic entity mentioned in questions, then traverse the knowledge graph to find relations as a multi-hop path to answers, while we propose a novel approach to leverage simple-question answerers to answer compound questions. Our model consists of two parts: (i) a novel learning-to-decompose agent that learns a policy to decompose a compound question into simple questions and (ii) three independent simple-question answerers that classify the corresponding relations for each simple question. Experiments demonstrate that our model learns complex rules of compositionality as stochastic policy, which benefits simple neural networks to achieve state-of-the-art results on WebQuestions and MetaQA. We analyze the interpretable decomposition process as well as generated partitions.

\*\*\*\*\*

#### Unsupervised Neural Multi-Document Abstractive Summarization of Reviews

Eric Chu, Peter J. Liu

Abstractive summarization has been studied using neural sequence transduction methods with datasets of large, paired document-summary examples. However, such datasets are rare and the models trained from them do not generalize to other domains. Recently, some progress has been made in learning sequence-to-sequence mappings with only unpaired examples. In our work, we consider the setting where there are only documents (product or business reviews) with no summaries provided, and propose an end-to-end, neural model architecture to perform unsupervised abstractive summarization. Our proposed model consists of an auto-encoder trained so that the mean of the representations of the input reviews decodes to a reasonable summary-review. We consider variants of the proposed architecture and perform an ablation study to show the importance of specific components. We show through metrics and human evaluation that the generated summaries are highly abstractive, fluent, relevant, and representative of the average sentiment of the input reviews.

\*\*\*\*\*

Learning Joint Wasserstein Auto-Encoders for Joint Distribution Matching  
Jiezhong Cao, Yong Guo, Langyuan Mo, Peilin Zhao, Junzhou Huang, Minghui Tan

We study the joint distribution matching problem which aims at learning bidirectional mappings to match the joint distribution of two domains. This problem occurs in unsupervised image-to-image translation and video-to-video synthesis tasks, which, however, has two critical challenges: (i) it is difficult to exploit sufficient information from the joint distribution; (ii) how to theoretically and experimentally evaluate the generalization performance remains an open question.

To address the above challenges, we propose a new optimization problem and design a novel Joint Wasserstein Auto-Encoders (JWAE) to minimize the Wasserstein distance of the joint distributions in two domains. We theoretically prove that the generalization ability of the proposed method can be guaranteed by minimizing the Wasserstein distance of joint distributions. To verify the generalization ability, we apply our method to unsupervised video-to-video synthesis by performing video frame interpolation and producing visually smooth videos in two domains, simultaneously. Both qualitative and quantitative comparisons demonstrate the superiority of our method over several state-of-the-arts.

\*\*\*\*\*

A Convergent Variant of the Boltzmann Softmax Operator in Reinforcement Learning  
Ling Pan, Qingpeng Cai, Qi Meng, Wei Chen, Tie-Yan Liu

The Boltzmann softmax operator can trade-off well between exploration and exploitation according to current estimation in an exponential weighting scheme, which is a promising way to address the exploration-exploitation dilemma in reinforcement learning. Unfortunately, the Boltzmann softmax operator is not a non-expansion, which may lead to unstable or even divergent learning behavior when used in estimating the value function. The non-expansion is a vital and widely-used sufficient condition to guarantee the convergence of value iteration. However, how to characterize the effect of such non-expansive operators in value iteration remains an open problem. In this paper, we propose a new technique to analyze the error bound of value iteration with the Boltzmann softmax operator. We then propose the dynamic Boltzmann softmax (DBS) operator to enable the convergence to the optimal value function in value iteration. We also present convergence rate analysis of the algorithm.

Using Q-learning as an application, we show that the DBS operator can be applied in a model-free reinforcement learning algorithm. Finally, we demonstrate the effectiveness of the DBS operator in a toy problem called GridWorld and a suite of Atari games. Experimental results show that outperforms DQN substantially in benchmark games.

\*\*\*\*\*

Recycling the discriminator for improving the inference mapping of GAN  
Duhyeon Bang, Hyunjung Shim

Generative adversarial networks (GANs) have achieved outstanding success in generating the high-quality data. Focusing on the generation process, existing GANs learn a unidirectional mapping from the latent vector to the data. Later, various studies point out that the latent space of GANs is semantically meaningful and can be utilized in advanced data analysis and manipulation. In order to analyze the real data in the latent space of GANs, it is necessary to investigate the inverse generation mapping from the data to the latent vector. To tackle this problem, the bidirectional generative models introduce an encoder to establish the inverse path of the generation process. Unfortunately, this effort leads to the degradation of generation quality because the imperfect generator rather interferes the encoder training and vice versa.

In this paper, we propose an effective algorithm to infer the latent vector based on existing unidirectional GANs by preserving their generation quality.

It is important to note that we focus on increasing the accuracy and efficiency of the inference mapping but not influencing the GAN performance (i.e., the quality or the diversity of the generated sample).

Furthermore, utilizing the proposed inference mapping algorithm, we suggest a new metric for evaluating the GAN models by measuring the reconstruction error of

unseen real data.

The experimental analysis demonstrates that the proposed algorithm achieves more accurate inference mapping than the existing method and provides the robust metric for evaluating GAN performance.

\*\*\*\*\*

#### Differentiable Expected BLEU for Text Generation

Wentao Wang,Zhiting Hu,Zichao Yang,Haoran Shi,Eric P. Xing

Neural text generation models such as recurrent networks are typically trained by maximizing data log-likelihood based on cross entropy. Such training objective shows a discrepancy from test criteria like the BLEU metric. Recent work optimizes expected BLEU under the model distribution using policy gradient, while such algorithm can suffer from high variance and become impractical. In this paper, we propose a new Differentiable Expected BLEU (DEBLEU) objective that permits direct optimization of neural generation models with gradient descent. We leverage the decomposability and sparsity of BLEU, and reformulate it with moderate approximations, making the evaluation of the objective and its gradient efficient, comparable to common cross-entropy loss. We further devise a simple training procedure with ground-truth masking and annealing for stable optimization. Experiments on neural machine translation and image captioning show our method significantly improves over both cross-entropy and policy gradient training.

\*\*\*\*\*

#### Learning Two-layer Neural Networks with Symmetric Inputs

Rong Ge,Rohith Kuditipudi,Zhize Li,Xiang Wang

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

$$y = A \sum Wx + \xi,$$

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

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.

\*\*\*\*\*

#### Meta-Learning to Guide Segmentation

Kate Rakelly\*,Evan Shelhamer\*,Trevor Darrell,Alexei A. Efros,Sergey Levine

There are myriad kinds of segmentation, and ultimately the "right" segmentation of a given scene is in the eye of the annotator. Standard approaches require large amounts of labeled data to learn just one particular kind of segmentation. As a first step towards relieving this annotation burden, we propose the problem of guided segmentation: given varying amounts of pixel-wise labels, segment unannotated pixels by propagating supervision locally (within an image) and non-locally (across images). We propose guided networks, which extract a latent task representation---guidance---from variable amounts and classes (categories, instances, etc.) of pixel supervision and optimize our architecture end-to-end for fast, accurate, and data-efficient segmentation by meta-learning. To span the few-shot and many-shot learning regimes, we examine guidance from as little as one pixel per concept to as much as 1000+ images, and compare to full gradient optimization at both extremes. To explore generalization, we analyze guidance as a bridge between different levels of supervision to segment classes as the union of instances. Our segmentor concentrates different amounts of supervision of different types of classes into an efficient latent representation, non-locally propagates this supervision across images, and can be updated quickly and cumulatively when given more supervision.

\*\*\*\*\*

Towards More Theoretically-Grounded Particle Optimization Sampling for Deep Learning

Jianyi Zhang, Ruiyi Zhang, Changyou Chen

Many deep-learning based methods such as Bayesian deep learning (DL) and deep reinforcement learning (RL) have heavily relied on the ability of a model being able to efficiently explore via Bayesian sampling. Particle-optimization sampling (POS) is a recently developed technique to generate high-quality samples from a target distribution by iteratively updating a set of interactive particles, with a representative algorithm the Stein variational gradient descent (SVGD). Though obtaining significant empirical success, the non-asymptotic convergence behavior of SVGD remains unknown. In this paper, we generalize POS to a stochasticity setting by injecting random noise in particle updates, called stochastic particle-optimization sampling (SPOS). Notably, for the first time, we develop non-asymptotic convergence theory for the SPOS framework, characterizing convergence of a sample approximation w.r.t. the number of particles and iterations under both convex- and nonconvex-energy-function settings. Interestingly, we provide theoretical understanding of a pitfall of SVGD that can be avoided in the proposed SPOS framework, i.e., particles tend to collapse to a local mode in SVGD under some particular conditions. Our theory is based on the analysis of nonlinear stochastic differential equations, which serves as an extension and a complementary development to the asymptotic convergence theory for SVGD such as (Liu, 2017). With such theoretical guarantees, SPOS can be safely and effectively applied on both Bayesian DL and deep RL tasks. Extensive results demonstrate the effectiveness of our proposed framework.

\*\*\*\*\*

SOLAR: Deep Structured Representations for Model-Based Reinforcement Learning

Marvin Zhang\*, Sharad Vikram\*, Laura Smith, Pieter Abbeel, Matthew Johnson, Sergey Levine

Model-based reinforcement learning (RL) methods can be broadly categorized as global model methods, which depend on learning models that provide sensible predictions in a wide range of states, or local model methods, which iteratively refit simple models that are used for policy improvement. While predicting future states that will result from the current actions is difficult, local model methods only attempt to understand system dynamics in the neighborhood of the current policy, making it possible to produce local improvements without ever learning to predict accurately far into the future. The main idea in this paper is that we can learn representations that make it easy to retrospectively infer simple dynamics given the data from the current policy, thus enabling local models to be used for policy learning in complex systems. We evaluate our approach against other model-based and model-free RL methods on a suite of robotics tasks, including manipulation tasks on a real Sawyer robotic arm directly from camera images.



\*\*\*\*\*

#### Fast Binary Functional Search on Graph

Shulong Tan,Zhixin Zhou,Zhaozhuo Xu,Ping Li

The large-scale search is an essential task in modern information systems. Numerous learning based models are proposed to capture semantic level similarity measures for searching or ranking. However, these measures are usually complicated and beyond metric distances. As Approximate Nearest Neighbor Search (ANNS) techniques have specifications on metric distances, efficient searching by advanced measures is still an open question. In this paper, we formulate large-scale search as a general task, Optimal Binary Functional Search (OBFS), which contains ANNS as special cases. We analyze existing OBFS methods' limitations and explain they are not applicable for complicated searching measures. We propose a flexible graph-based solution for OBFS, Search on L2 Graph (SL2G). SL2G approximates gradient decent in Euclidean space, with accessible conditions. Experiments demonstrate SL2G's efficiency in searching by advanced matching measures (i.e., Neural Network based measures).

\*\*\*\*\*

#### Spectral Inference Networks: Unifying Deep and Spectral Learning

David Pfau,Stig Petersen,Ashish Agarwal,David G. T. Barrett,Kimberly L. Stachenfeld

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

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.

\*\*\*\*\*

#### ACE: Artificial Checkerboard Enhancer to Induce and Evade Adversarial Attacks

Jisung Hwang,Younghoon Kim,Sanghyuk Chun,Jaejun Yoo, Ji-Hoon Kim,Dongyoon Han,Jun-g-Woo Ha

The checkerboard phenomenon is one of the well-known visual artifacts in the computer vision field. The origins and solutions of checkerboard artifacts in the pixel space have been studied for a long time, but their effects on the gradient space have rarely been investigated. In this paper, we revisit the checkerboard artifacts in the gradient space which turn out to be the weak point of a network architecture. We explore image-agnostic property of gradient checkerboard artifacts and propose a simple yet effective defense method by utilizing the artifacts. We introduce our defense module, dubbed Artificial Checkerboard Enhancer (ACE), which induces adversarial attacks on designated pixels. This enables the model to deflect attacks by shifting only a single pixel in the image with a remarkable

ble defense rate. We provide extensive experiments to support the effectiveness of our work for various attack scenarios using state-of-the-art attack methods. Furthermore, we show that ACE is even applicable to large-scale datasets including ImageNet dataset and can be easily transferred to various pretrained networks

\*\*\*\*\*

Co-manifold learning with missing data

Gal Mishne, Eric C. Chi, Ronald R. Coifman

Representation learning is typically applied to only one mode of a data matrix, either its rows or columns. Yet in many applications, there is an underlying geometry to both the rows and the columns. We propose utilizing this coupled structure to perform co-manifold learning: uncovering the underlying geometry of both the rows and the columns of a given matrix, where we focus on a missing data setting. Our unsupervised approach consists of three components. We first solve a family of optimization problems to estimate a complete matrix at multiple scales of smoothness. We then use this collection of smooth matrix estimates to compute pairwise distances on the rows and columns based on a new multi-scale metric that implicitly introduces a coupling between the rows and the columns. Finally, we construct row and column representations from these multi-scale metrics. We demonstrate that our approach outperforms competing methods in both data visualization and clustering.

\*\*\*\*\*

Unification of Recurrent Neural Network Architectures and Quantum Inspired Stable Design

Murphy Yuezhen Niu, Lior Horesh, Michael O'Keefe, Isaac Chuang

Various architectural advancements in the design of recurrent neural networks (RNN) have been focusing on improving the empirical stability and representability by sacrificing the complexity of the architecture. However, more remains to be done to fully understand the fundamental trade-off between these conflicting requirements. Towards answering this question, we forsake the purely bottom-up approach of data-driven machine learning to understand, instead, the physical origin and dynamical properties of existing RNN architectures. This facilitates designing new RNNs with smaller complexity overhead and provable stability guarantee. First, we define a family of deep recurrent neural networks,  $\mathcal{RNN}_{\sigma, \tau}$ , according to the order of nonlinearity  $\sigma$  and the range of temporal memory scale  $\tau$  in their underlying dynamics embodied in the form of discretized ordinary differential equations. We show that most of the existing proposals of RNN architectures belong to different orders of  $\mathcal{RNN}_{\sigma, \tau}$ . We then propose a new RNN ansatz, namely the Quantum-inspired Universal computing Neural Network (QUNN), to leverage the reversibility, stability, and universality of quantum computation for stable and universal RNN. QUNN provides a complexity reduction in the number of training parameters from being polynomial in both data and correlation time to only linear in correlation time. Compared to Long-Short-Term Memory (LSTM), QUNN of the same number of hidden layers facilitates higher nonlinearity and longer memory span with provable stability. Our work opens new directions in designing minimal RNNs based on additional knowledge about the dynamical nature of both the data and different training architectures.

\*\*\*\*\*

Aligning Artificial Neural Networks to the Brain yields Shallow Recurrent Architectures

Jonas Kubilius, Martin Schrimpf, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, James J. DiCarlo

Deep artificial neural networks with spatially repeated processing (a.k.a., deep convolutional ANNs) have been established as the best class of candidate models of visual processing in the primate ventral visual processing stream. Over the past five years, these ANNs have evolved from a simple feedforward eight-layer architecture in AlexNet to extremely deep and branching NASNet architectures, demonstrating increasingly better object categorization performance. Here we ask, as ANNs have continued to evolve in performance, are they also strong candidate m

models for the brain? To answer this question, we developed Brain-Score, a composite of neural and behavioral benchmarks for determining how brain-like a model is, together with an online platform where models can receive a Brain-Score and compare against other models.

Despite high scores, typical deep models from the machine learning community are often hard to map onto the brain's anatomy due to their vast number of layers and missing biologically-important connections, such as recurrence. To further map onto anatomy and validate our approach, we built CORnet-S: an ANN guided by Brain-Score with the anatomical constraints of compactness and recurrence. Although a shallow model with four anatomically mapped areas and recurrent connectivity, CORnet-S is a top model on Brain-Score and outperforms similarly compact models on ImageNet. Analyzing CORnet-S circuitry variants revealed recurrence as the main predictive factor of both Brain-Score and ImageNet top-1 performance.

\*\*\*\*\*

#### Selective Convolutional Units: Improving CNNs via Channel Selectivity

Jongheon Jeong, Jinwoo Shin

Bottleneck structures with identity (e.g., residual) connection are now emerging popular paradigms for designing deep convolutional neural networks (CNN), for processing large-scale features efficiently. In this paper, we focus on the information-preserving nature of identity connection and utilize this to enable a convolutional layer to have a new functionality of channel-selectivity, i.e., redistributing its computations to important channels. In particular, we propose Selective Convolutional Unit (SCU), a widely-applicable architectural unit that improves parameter efficiency of various modern CNNs with bottlenecks. During training, SCU gradually learns the channel-selectivity on-the-fly via the alternative usage of (a) pruning unimportant channels, and (b) rewiring the pruned parameters to important channels. The rewired parameters emphasize the target channel in a way that selectively enlarges the convolutional kernels corresponding to it. Our experimental results demonstrate that the SCU-based models without any postprocessing generally achieve both model compression and accuracy improvement compared to the baselines, consistently for all tested architectures.

\*\*\*\*\*

#### AutoLoss: Learning Discrete Schedule for Alternate Optimization

Haowen Xu, Hao Zhang, Zhiting Hu, Xiaodan Liang, Ruslan Salakhutdinov, Eric Xing

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.

\*\*\*\*\*

#### Graph Classification with Geometric Scattering

Feng Gao, Guy Wolf, Matthew Hirn

One of the most notable contributions of deep learning is the application of convolutional neural networks (ConvNets) to structured signal classification, and in particular image classification. Beyond their impressive performances in supervised learning, the structure of such networks inspired the development of deep filter banks referred to as scattering transforms. These transforms apply a cascade

ade of wavelet transforms and complex modulus operators to extract features that are invariant to group operations and stable to deformations. Furthermore, Conv Nets inspired recent advances in geometric deep learning, which aim to generalize these networks to graph data by applying notions from graph signal processing to learn deep graph filter cascades. We further advance these lines of research by proposing a geometric scattering transform using graph wavelets defined in terms of random walks on the graph. We demonstrate the utility of features extracted with this designed deep filter bank in graph classification of biochemistry and social network data (incl. state of the art results in the latter case), and in data exploration, where they enable inference of EC exchange preferences in enzyme evolution.

\*\*\*\*\*

Deli-Fisher GAN: Stable and Efficient Image Generation With Structured Latent Generative Space

Boli Fang, Chuck Jia, Miao Jiang, Dhawal Chaturvedi

Generative Adversarial Networks (GANs) are powerful tools for realistic image generation. However, a major drawback of GANs is that they are especially hard to train, often requiring large amounts of data and long training time. In this paper we propose the Deli-Fisher GAN, a GAN that generates photo-realistic images by enforcing structure on the latent generative space using similar approaches in \cite{deligan}. The structure of the latent space we consider in this paper is modeled as a mixture of Gaussians, whose parameters are learned in the training process. Furthermore, to improve stability and efficiency, we use the Fisher Integral Probability Metric as the divergence measure in our GAN model, instead of the Jensen-Shannon divergence. We show by experiments that the Deli-Fisher GAN performs better than DCGAN, WGAN, and the Fisher GAN as measured by inception score.

\*\*\*\*\*

An Energy-Based Framework for Arbitrary Label Noise Correction

Jaspreet Sahota, Divya Shanmugam, Janahan Ramanan, Sepehr Eghbali, Marcus Brubaker

We propose an energy-based framework for correcting mislabelled training examples in the context of binary classification. While existing work addresses random and class-dependent label noise, we focus on feature dependent label noise, which is ubiquitous in real-world data and difficult to model. Two elements distinguish our approach from others: 1) instead of relying on the original feature space, we employ an autoencoder to learn a discriminative representation and 2) we introduce an energy-based formalism for the label correction problem. We prove that a discriminative representation can be learned by training a generative model using a loss function comprised of the difference of energies corresponding to each class. The learned energy value for each training instance is compared to the original training labels and contradictions between energy assignment and training label are used to correct labels. We validate our method across eight datasets, spanning synthetic and realistic settings, and demonstrate the technique's state-of-the-art label correction performance. Furthermore, we derive analytical expressions to show the effect of label noise on the gradients of empirical risk.

\*\*\*\*\*

Analyzing Inverse Problems with Invertible Neural Networks

Lynton Ardizzone, Jakob Kruse, Carsten Rother, Ullrich Köthe

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.

\*\*\*\*\*

The wisdom of the crowd: reliable deep reinforcement learning through ensembles of Q-functions

Daniel Elliott, Charles Anderson

Reinforcement learning agents learn by exploring the environment and then exploiting what they have learned.

This frees the human trainers from having to know the preferred action or intrinsic value of each encountered state.

The cost of this freedom is reinforcement learning is slower and more unstable than supervised learning.

We explore the possibility that ensemble methods can remedy these shortcomings and do so by investigating a novel technique which harnesses the wisdom of the crowds by bagging Q-function approximator estimates.

Our results show that this proposed approach improves all three tasks and reinforcement learning approaches attempted.

We are able to demonstrate that this is a direct result of the increased stability of the action portion of the state-action-value function used by Q-learning to select actions and by policy gradient methods to train the policy.

\*\*\*\*\*

Learning Finite State Representations of Recurrent Policy Networks

Anurag Koul, Alan Fern, Sam Greisdan

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.

\*\*\*\*\*

Dynamic Planning Networks

Norman L. Tasfi, Miriam Capretz

We introduce Dynamic Planning Networks (DPN), a novel architecture for deep reinforcement learning, that combines model-based and model-free aspects for online planning. Our architecture learns to dynamically construct plans using a learned state-transition model by selecting and traversing between simulated states and actions to maximize valuable information before acting. In contrast to model-free methods, model-based planning lets the agent efficiently test action hypotheses without performing costly trial-and-error in the environment. DPN learns to efficiently form plans by expanding a single action-conditional state transition at a time instead of exhaustively evaluating each action, reducing the required number of state-transitions during planning by up to 96%. We observe various emergent planning patterns used to solve environments, including classical search methods such as breadth-first and depth-first search. Learning To Plan shows improved data efficiency, performance, and generalization to new and unseen domains in comparison to several baselines.

\*\*\*\*\*

## Automata Guided Skill Composition

Xiao Li, Yao Ma, Calin Belta

Skills learned through (deep) reinforcement learning often generalize poorly across tasks and re-training is necessary when presented with a new task. We present a framework that combines techniques in formal methods with reinforcement

learning (RL) that allows for the convenient specification of complex temporal dependent tasks with logical expressions and construction of new skills from existing

ones with no additional exploration. We provide theoretical results for our composition technique and evaluate on a simple grid world simulation as well as a robotic manipulation task.

\*\*\*\*\*

## Measuring and regularizing networks in function space

Ari Benjamin, David Rolnick, Konrad Kording

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

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.

\*\*\*\*\*

## N-Ary Quantization for CNN Model Compression and Inference Acceleration

Günther Schindler, Wolfgang Roth, Franz Pernkopf, Holger Fröning

The tremendous memory and computational complexity of Convolutional Neural Networks (CNNs) prevents the inference deployment on resource-constrained systems. As a result, recent research focused on CNN optimization techniques, in particular quantization, which allows weights and activations of layers to be represented with just a few bits while achieving impressive prediction performance. However, aggressive quantization techniques still fail to achieve full-precision prediction performance on state-of-the-art CNN architectures on large-scale classification tasks. In this work we propose a method for weight and activation quantization that is scalable in terms of quantization levels (n-ary representations) and easy to compute while maintaining the performance close to full-precision CNNs. Our weight quantization scheme is based on trainable scaling factors and a nested

d-means clustering strategy which is robust to weight updates and therefore exhibits good convergence properties. The flexibility of nested-means clustering enables exploration of various n-ary weight representations with the potential of high parameter compression. For activations, we propose a linear quantization strategy that takes the statistical properties of batch normalization into account. We demonstrate the effectiveness of our approach using state-of-the-art models on ImageNet.

\*\*\*\*\*

#### Knowledge Distillation from Few Samples

Tianhong Li, Jianguo Li, Zhuang Liu, Changshui Zhang

Current knowledge distillation methods require full training data to distill knowledge from a large "teacher" network to a compact "student" network by matching certain statistics between "teacher" and "student" such as softmax outputs and feature responses. This is not only time-consuming but also inconsistent with human cognition in which children can learn knowledge from adults with few examples. This paper proposes a novel and simple method for knowledge distillation from few samples. Taking the assumption that both "teacher" and "student" have the same feature map sizes at each corresponding block, we add a  $1 \times 1$  conv-layer at the end of each block in the student-net, and align the block-level outputs between "teacher" and "student" by estimating the parameters of the added layer with limited samples. We prove that the added layer can be absorbed/merged into the previous conv-layer to formulate a new conv-layer with the same size of parameters and computation cost as previous one. Experiments verify that the proposed method is very efficient and effective to distill knowledge from teacher-net to student-net constructing in different ways on various datasets.

\*\*\*\*\*

#### Deep Frank-Wolfe For Neural Network Optimization

Leonard Berrada, Andrew Zisserman, M. Pawan Kumar

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 BPGGrad 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>.

\*\*\*\*\*

#### A Deep Learning Approach for Dynamic Survival Analysis with Competing Risks

Changhee Lee, Mihaela van der Schaar

Currently available survival analysis methods are limited in their ability to deal with complex, heterogeneous, and longitudinal data such as that available in primary care records, or in their ability to deal with multiple competing risks.

This paper develops a novel deep learning architecture that flexibly incorporates the available longitudinal data comprising various repeated measurements (rather than only the last available measurements) in order to issue dynamically updated survival predictions for one or multiple competing risk(s). Unlike existing works in the survival analysis on the basis of longitudinal data, the proposed

method learns the time-to-event distributions without specifying underlying stochastic assumptions of the longitudinal or the time-to-event processes. Thus, our method is able to learn associations between the longitudinal data and the various associated risks in a fully data-driven fashion. We demonstrate the power of our method by applying it to real-world longitudinal datasets and show a drastic improvement over state-of-the-art methods in discriminative performance. Furthermore, our analysis of the variable importance and dynamic survival predictions will yield a better understanding of the predicted risks which will result in more effective health care.

\*\*\*\*\*

Quasi-hyperbolic momentum and Adam for deep learning

Jerry Ma, Denis Yarats

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.

\*\*\*\*\*

LIT: Block-wise Intermediate Representation Training for Model Compression

Animesh Koratana\*, Daniel Kang\*, Peter Bailis, Matei Zaharia

Knowledge distillation (KD) is a popular method for reducing the computational overhead of deep network inference, in which the output of a teacher model is used to train

a smaller, faster student model. Hint training (i.e., FitNets) extends KD by regressing a

student model's intermediate representation to a teacher model's intermediate representation.

In this work, we introduce block-wise Intermediate representation Training (LIT),

a novel model compression technique that extends the use of intermediate representations in deep network compression, outperforming KD and hint training. LIT has two

key ideas: 1) LIT trains a student of the same width (but shallower depth) as the teacher

by directly comparing the intermediate representations, and 2) LIT uses the intermediate

representation from the previous block in the teacher model as an input to the current student

block during training, avoiding unstable intermediate representations in the student

network. We show that LIT provides substantial reductions in network depth without

loss in accuracy – for example, LIT can compress a ResNeXt-110 to a ResNeXt-20 (5.5×) on CIFAR10 and a VDCNN-29 to a VDCNN-9 (3.2×) on Amazon Reviews

without loss in accuracy, outperforming KD and hint training in network size at a given

accuracy. We also show that applying LIT to identical student/teacher architectures

increases the accuracy of the student model above the teacher model, outperforming the

recently-proposed Born Again Networks procedure on ResNet, ResNeXt, and VDCNN. Finally, we show that LIT can effectively compress GAN generators.



\*\*\*\*\*

## REPRESENTATION COMPRESSION AND GENERALIZATION IN DEEP NEURAL NETWORKS

Ravid Shwartz-Ziv, Amichai Painsky, Naftali Tishby

Understanding the groundbreaking performance of Deep Neural Networks is one of the greatest challenges to the scientific community today. In this work, we introduce an information theoretic viewpoint on the behavior of deep networks optimization processes and their generalization abilities. By studying the Information

Plane, the plane of the mutual information between the input variable and the desired label, for each hidden layer. Specifically, we show that the training of

the network is characterized by a rapid increase in the mutual information (MI) between the layers and the target label, followed by a longer decrease in the MI between the layers and the input variable. Further, we explicitly show that these

two fundamental information-theoretic quantities correspond to the generalization

error of the network, as a result of introducing a new generalization bound that is

exponential in the representation compression. The analysis focuses on typical patterns of large-scale problems. For this purpose, we introduce a novel analytic

bound on the mutual information between consecutive layers in the network.

An important consequence of our analysis is a super-linear boost in training time

with the number of non-degenerate hidden layers, demonstrating the computational benefit of the hidden layers.

\*\*\*\*\*

## Adversarially Robust Training through Structured Gradient Regularization

Kevin Roth, Aurelien Lucchi, Sebastian Nowozin, Thomas Hofmann

We propose a novel data-dependent structured gradient regularizer to increase the robustness of neural networks vis-a-vis adversarial perturbations. Our regularizer can be derived as a controlled approximation from first principles, leveraging the fundamental link between training with noise and regularization. It adds very little computational overhead during learning and is simple to implement generically in standard deep learning frameworks. Our experiments provide strong evidence that structured gradient regularization can act as an effective first line of defense against attacks based on long-range correlated signal corruptions.

\*\*\*\*\*

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

Haoming Jiang, Zhehui Chen, Minshuo Chen, Feng Liu, Dingding Wang, Tuo Zhao

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 ImgaNet 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 S

## Speech Recognition

Chun-Fu (Richard) Chen, Quanfu Fan, Neil Mallinar, Tom Sercu, Rogerio Feris

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.

\*\*\*\*\*

## An adaptive homeostatic algorithm for the unsupervised learning of visual features

Victor Boutin, Angelo Franciosini, Laurent Perrinet

The formation of structure in the brain, that is, of the connections between cells within neural populations, is by large an unsupervised learning process: the emergence of this architecture is mostly self-organized. In the primary visual cortex of mammals, for example, one may observe during development the formation of cells selective to localized, oriented features. This leads to the development of a rough representation of contours of the retinal image in area V1. We modeled these mechanisms using sparse Hebbian learning algorithms. These algorithms alternate a coding step to encode the information with a learning step to find the proper encoder. A major difficulty faced by these algorithms is to deduce a good representation while knowing immature encoders, and to learn good encoders with a non-optimal representation. To address this problem, we propose to introduce a new regulation process between learning and coding, called homeostasis. Our homeostasis is compatible with a neuro-mimetic architecture and allows for the fast emergence of localized filters sensitive to orientation. The key to this algorithm lies in a simple adaptation mechanism based on non-linear functions that reconciles the antagonistic processes that occur at the coding and learning time scales. We tested this unsupervised algorithm with this homeostasis rule for a range of existing unsupervised learning algorithms coupled with different neural coding algorithms. In addition, we propose a simplification of this optimal homeostasis rule by implementing a simple heuristic on the probability of activation of neurons. Compared to the optimal homeostasis rule, we show that this heuristic allows to implement a more rapid unsupervised learning algorithm while keeping a large part of its effectiveness. These results demonstrate the potential application of such a strategy in machine learning and we illustrate this with one result in a convolutional neural network.

\*\*\*\*\*

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

Michael Lutter, Christian Ritter, Jan Peters

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 e

quations 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.

\*\*\*\*\*

Learning Disentangled Representations with Reference-Based Variational Autoencoders

Adria Ruiz, Oriol Martinez, Xavier Binefa, Jakob Verbeek

Learning disentangled representations from visual data, where different high-level generative factors are independently encoded, is of importance for many computer vision tasks. Supervised approaches, however, require a significant annotation effort in order to label the factors of interest in a training set. To alleviate the annotation cost, we introduce a learning setting which we refer to as "reference-based disentangling". Given a pool of unlabelled images, the goal is to learn a representation where a set of target factors are disentangled from others. The only supervision comes from an auxiliary "reference set" that contains images where the factors of interest are constant. In order to address this problem, we propose reference-based variational autoencoders, a novel deep generative model designed to exploit the weak supervisory signal provided by the reference set. During training, we use the variational inference framework where adversarial learning is used to minimize the objective function. By addressing tasks such as feature learning, conditional image generation or attribute transfer, we validate the ability of the proposed model to learn disentangled representations from minimal supervision.

\*\*\*\*\*

Adversarial Audio Synthesis

Chris Donahue, Julian McAuley, Miller Puckette

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

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.

\*\*\*\*\*

Interpreting Adversarial Robustness: A View from Decision Surface in Input Space  
Fuxun Yu, Chenchen Liu, Yanzhi Wang, Xiang Chen

One popular hypothesis of neural network generalization is that the flat local minima of loss surface in parameter space leads to good generalization. However, we demonstrate that loss surface in parameter space has no obvious relationship with generalization, especially under adversarial settings. Through visualizing decision surfaces in both parameter space and input space, we instead show that the geometry property of decision surface in input space correlates well with the adversarial robustness. We then propose an adversarial robustness indicator, which can evaluate a neural network's intrinsic robustness property without testing its accuracy under adversarial attacks. Guided by it, we further propose our robust training method. Without involving adversarial training, our method could enhance network's intrinsic adversarial robustness against various adversarial attacks.

\*\*\*\*\*

The Laplacian in RL: Learning Representations with Efficient Approximations  
Yifan Wu, George Tucker, Ofir Nachum

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 Jastrzebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, Amos Storkey

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.

\*\*\*\*\*

## Detecting Topological Defects in 2D Active Nematics Using Convolutional Neural Networks

Ruoshi Liu, Michael M. Norton, Seth Fraden, Pengyu Hong

Active matter consists of active agents which transform energy extracted from surroundings into momentum, producing a variety of collective phenomena. A model, synthetic active system composed of microtubule polymers driven by protein motors spontaneously forms a liquid-crystalline nematic phase. Extensile stress created by the protein motors precipitates continuous buckling and folding of the microtubules creating motile topological defects and turbulent fluid flows. Defect motion is determined by the rheological properties of the material; however, these remain largely unquantified. Measuring defects dynamics can yield fundamental insights into active nematics, a class of materials that include bacterial films and animal cells. Current methods for defect detection lack robustness and precision, and require fine-tuning for datasets with different visual quality. In this study, we applied Deep Learning to train a defect detector to automatically analyze microscopy videos of the microtubule active nematic. Experimental results indicate that our method is robust and accurate. It is expected to significantly increase the amount of video data that can be processed.

\*\*\*\*\*

## Constrained Bayesian Optimization for Automatic Chemical Design

Ryan-Rhys Griffiths, José Miguel Hernández-Lobato

Automatic Chemical Design provides a framework for generating novel molecules with optimized molecular properties. The current model suffers from the pathology that it tends to produce invalid molecular structures. By reformulating the search procedure as a constrained Bayesian optimization problem, we showcase improvements in both the validity and quality of the generated molecules. We demonstrate that the model consistently produces novel molecules ranking above the 90th percentile of the distribution over training set scores across a range of objective functions. Importantly, our method suffers no degradation in the complexity or the diversity of the generated molecules.

\*\*\*\*\*

## ChoiceNet: Robust Learning by Revealing Output Correlations

Sungjoon Choi, Sanghoon Hong, Kyungjae Lee, Sungbin Lim

In this paper, we focus on the supervised learning problem with corrupt training data. We assume that the training dataset is generated from a mixture of a target distribution and other unknown distributions. We estimate the quality of each data by revealing the correlation between the generated distribution and the target distribution. To this end, we present a novel framework referred to here as ChoiceNet that can robustly infer the target distribution in the presence of inconsistent data. We demonstrate that the proposed framework is applicable to both classification and regression tasks. Particularly, ChoiceNet is evaluated in comprehensive experiments, where we show that it constantly outperforms existing baseline methods in the handling of noisy data in synthetic regression tasks as well as behavior cloning problems. In the classification tasks, we apply the proposed method to the MNIST and CIFAR-10 datasets and it shows superior performances in terms of robustness to different types of noisy labels.

\*\*\*\*\*

## Graph2Seq: Scalable Learning Dynamics for Graphs

Shaileshh Bojja Venkatakrishnan, Mohammad Alizadeh, Pramod Viswanath

Neural networks have been shown to be an effective tool for learning algorithms over graph-structured data. However, graph representation techniques---that convert graphs to real-valued vectors for use with neural networks---are still in their infancy. Recent works have proposed several approaches (e.g., graph convolutional networks), but these methods have difficulty scaling and generalizing to graphs with different sizes and shapes. We present Graph2Seq, a new technique that represents vertices of graphs as infinite time-series. By not limiting the representation to a fixed dimension, Graph2Seq scales naturally to graphs of arbitrary sizes and shapes. Graph2Seq is also reversible, allowing full recovery of the graph structure from the sequence. By analyzing a formal computational model f

or graph representation, we show that an unbounded sequence is necessary for scalability. Our experimental results with Graph2Seq show strong generalization and new state-of-the-art performance on a variety of graph combinatorial optimization problems.

\*\*\*\*\*

#### Learning Corresponded Rationales for Text Matching

Mo Yu, Shiyu Chang, Tommi S Jaakkola

The ability to predict matches between two sources of text has a number of applications including natural language inference (NLI) and question answering (QA). While flexible neural models have become effective tools in solving these tasks, they are rarely transparent in terms of the mechanism that mediates the prediction. In this paper, we propose a self-explaining architecture where the model is forced to highlight, in a dependent manner, how spans of one side of the input match corresponding segments of the other side in order to arrive at the overall decision. The text spans are regularized to be coherent and concise, and their correspondence is captured explicitly. The text spans -- rationales -- are learned entirely as latent mechanisms, guided only by the distal supervision from the end-to-end task. We evaluate our model on both NLI and QA using three publicly available datasets. Experimental results demonstrate quantitatively and qualitatively that our method delivers interpretable justification of the prediction without sacrificing state-of-the-art performance. Our code and data split will be publicly available.

\*\*\*\*\*

#### TherML: The Thermodynamics of Machine Learning

Alexander A. Alemi, Ian Fischer

In this work we offer an information-theoretic framework for representation learning that connects with a wide class of existing objectives in machine learning. We develop a formal correspondence between this work and thermodynamics and discuss its implications.

\*\*\*\*\*

#### Geomstats: a Python Package for Riemannian Geometry in Machine Learning

Nina Miolane, Johan Mathe, Claire Donnat, Mikael Jorda, Xavier Pennec

We introduce geomstats, a Python package for Riemannian modelization and optimization over manifolds such as hyperspheres, hyperbolic spaces, SPD matrices or Lie groups of transformations. Our contribution is threefold. First, geomstats allows the flexible modeling of many a machine learning problem through an efficient and extensively unit-tested implementations of these manifolds, as well as the set of useful Riemannian metrics, exponential and logarithm maps that we provide. Moreover, the wide choice of loss functions and our implementation of the corresponding gradients allow fast and easy optimization over manifolds. Finally, geomstats is the only package to provide a unified framework for Riemannian geometry, as the operations implemented in geomstats are available with different computing backends (numpy, tensorflow and keras), as well as with a GPU-enabled mode --thus considerably facilitating the application of Riemannian geometry in machine learning. In this paper, we present geomstats through a review of the utility and advantages of manifolds in machine learning, using the concrete examples that they span to show the efficiency and practicality of their implementation using our package

\*\*\*\*\*

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

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, Jonathan Tompson

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

r 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.

\*\*\*\*\*

#### Generative Feature Matching Networks

Cicero Nogueira dos Santos, Inkit Padhi, Pierre Dognin, Youssef Mroueh

We propose a non-adversarial feature matching-based approach to train generative models. Our approach, Generative Feature Matching Networks (GFMN), leverages pretrained neural networks such as autoencoders and ConvNet classifiers to perform feature extraction. We perform an extensive number of experiments with different challenging datasets, including ImageNet. Our experimental results demonstrate that, due to the expressiveness of the features from pretrained ImageNet classifiers, even by just matching first order statistics, our approach can achieve state-of-the-art results for challenging benchmarks such as CIFAR10 and STL10.

\*\*\*\*\*

#### The GAN Landscape: Losses, Architectures, Regularization, and Normalization

Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, Sylvain Gelly

Generative adversarial networks (GANs) are a class of deep generative models which aim to learn a target distribution in an unsupervised fashion. While they were successfully applied to many problems, training a GAN is a notoriously challenging task and requires a significant amount of hyperparameter tuning, neural architecture engineering, and a non-trivial amount of "tricks". The success in many practical applications coupled with the lack of a measure to quantify the failure modes of GANs resulted in a plethora of proposed losses, regularization and normalization schemes, and neural architectures. In this work we take a sober view of the current state of GANs from a practical perspective. We reproduce the current state of the art and go beyond fairly exploring the GAN landscape. We discuss common pitfalls and reproducibility issues, open-source our code on Github, and provide pre-trained models on TensorFlow Hub.

\*\*\*\*\*

#### Hint-based Training for Non-Autoregressive Translation

Zhuohan Li, Di He, Fei Tian, Tao Qin, Liwei Wang, Tie-Yan Liu

Machine translation is an important real-world application, and neural network-based AutoRegressive Translation (ART) models have achieved very promising accuracy. Due to the unparallelizable nature of the autoregressive factorization, ART models have to generate tokens one by one during decoding and thus suffer from high inference latency. Recently, Non-AutoRegressive Translation (NART) models were proposed to reduce the inference time. However, they could only achieve inferior accuracy compared with ART models. To improve the accuracy of NART models, in this paper, we propose to leverage the hints from a well-trained ART model to train the NART model. We define two hints for the machine translation task: hints from hidden states and hints from word alignments, and use such hints to regularize the optimization of NART models. Experimental results show that the NART model trained with hints could achieve significantly better translation performance than previous NART models on several tasks. In particular, for the WMT14 En-De and De-En task, we obtain BLEU scores of 25.20 and 29.52 respectively, which largely outperforms the previous non-autoregressive baselines. It is even comparable to a strong LSTM-based ART model (24.60 on WMT14 En-De), but one order of magnitude faster in inference.

\*\*\*\*\*

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

Jacob Menick, Nal Kalchbrenner

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 CelebA HQ 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.

\*\*\*\*\*

What Information Does a ResNet Compress?

Luke Nicholas Darlow, Amos Storkey

The information bottleneck principle (Shwartz-Ziv & Tishby, 2017) suggests that SGD-based training of deep neural networks results in optimally compressed hidden layers, from an information theoretic perspective. However, this claim was established on toy data. The goal of the work we present here is to test these claims in a realistic setting using a larger and deeper convolutional architecture, a ResNet model. We trained PixelCNN++ models as inverse representation decoders to measure the mutual information between hidden layers of a ResNet and input image data, when trained for (1) classification and (2) autoencoding. We find that two stages of learning happen for both training regimes, and that compression does occur, even for an autoencoder. Sampling images by conditioning on hidden layers' activations offers an intuitive visualisation to understand what a ResNet learns to forget.

\*\*\*\*\*

Excessive Invariance Causes Adversarial Vulnerability

Joern-Henrik Jacobsen, Jens Behrmann, Richard Zemel, Matthias Bethge

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

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 va



riant 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.

\*\*\*\*\*

#### Guided Exploration in Deep Reinforcement Learning

Sahisnu Mazumder, Bing Liu, Shuai Wang, Yingxuan Zhu, Xiaotian Yin, Lifeng Liu, Jian Li, Yongbing Huang

This paper proposes a new method to drastically speed up deep reinforcement learning (deep RL) training for problems that have the property of  $\text{state-action permissibility}$  (SAP). Two types of permissibility are defined under SAP. The first type says that after an action  $a_t$  is performed in a state  $s_t$  and the agent reaches the new state  $s_{t+1}$ , the agent can decide whether the action  $a_t$  is  $\text{permissible}$  or  $\text{not permissible}$  in state  $s_t$ . The second type says that even without performing the action  $a_t$  in state  $s_t$ , the agent can already decide whether  $a_t$  is permissible or not in  $s_t$ . An action is not permissible in a state if the action can never lead to an optimal solution and thus should not be tried. We incorporate the proposed SAP property into two state-of-the-art deep RL algorithms to guide their state-action exploration. Results show that the SAP guidance can markedly speed up training.

\*\*\*\*\*

#### End-to-End Learning of Video Compression Using Spatio-Temporal Autoencoders

Jorge Pessoa, Helena Aidos, Pedro Tomás, Mário A. T. Figueiredo

Deep learning (DL) is having a revolutionary impact in image processing, with DL-based approaches now holding the state of the art in many tasks, including image compression. However, video compression has so far resisted the DL revolution, with the very few proposed approaches being based on complex and impractical architectures with multiple networks. This paper proposes what we believe is the first approach to end-to-end learning of a single network for video compression. We tackle the problem in a novel way, avoiding explicit motion estimation/prediction, by formalizing it as the rate-distortion optimization of a single spatio-temporal autoencoder; i.e., we jointly learn a latent-space projection transform and a synthesis transform for low bitrate video compression. The quantizer uses a rounding scheme, which is relaxed during training, and an entropy estimation technique to enforce an information bottleneck, inspired by recent advances in image compression. We compare the obtained video compression networks with standard widely-used codecs, showing better performance than the MPEG-4 standard, being competitive with H.264/AVC for low bitrates.

\*\*\*\*\*

#### Reliable Uncertainty Estimates in Deep Neural Networks using Noise Contrastive Priors

Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, James Davidson

Obtaining reliable uncertainty estimates of neural network predictions is a long standing challenge. Bayesian neural networks have been proposed as a solution, but it remains open how to specify their prior. In particular, the common practice of a standard normal prior in weight space imposes only weak regularities, causing the function posterior to possibly generalize in unforeseen ways on inputs outside of the training distribution. We propose noise contrastive priors (NCPs) to obtain reliable uncertainty estimates. The key idea is to train the model to output high uncertainty for data points outside of the training distribution.

NCPs do so using an input prior, which adds noise to the inputs of the current mini batch, and an output prior, which is a wide distribution given these inputs.

NCPs are compatible with any model that can output uncertainty estimates, are easy to scale, and yield reliable uncertainty estimates throughout training. Empirically, we show that NCPs prevent overfitting outside of the training distribution and result in uncertainty estimates that are useful for active learning. We demonstrate the scalability of our method on the flight delays data set, where we significantly improve upon previously published results.

\*\*\*\*\*

Padam: Closing the Generalization Gap of Adaptive Gradient Methods in Training Deep Neural Networks

Jinghui Chen, Quanquan Gu

Adaptive gradient methods, which adopt historical gradient information to automatically adjust the learning rate, despite the nice property of fast convergence, have been observed to generalize worse than stochastic gradient descent (SGD) with momentum in training deep neural networks. This leaves how to close the generalization gap of adaptive gradient methods an open problem. In this work, we show that adaptive gradient methods such as Adam, Amsgrad, are sometimes "over adapted". We design a new algorithm, called Partially adaptive momentum estimation method (Padam), which unifies the Adam/Amsgrad with SGD by introducing a partial adaptive parameter  $p$ , to achieve the best from both worlds. Experiments on standard benchmarks show that Padam can maintain fast convergence rate as Adam/Amsgrad while generalizing as well as SGD in training deep neural networks. These results would suggest practitioners pick up adaptive gradient methods once again for faster training of deep neural networks.

\*\*\*\*\*

AntisymmetricRNN: A Dynamical System View on Recurrent Neural Networks

Bo Chang, Minmin Chen, Eldad Haber, Ed H. Chi

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

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

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 our prior unsupervised learning methods.

\*\*\*\*\*

## Recurrent Experience Replay in Distributed Reinforcement Learning

Steven Kapturowski, Georg Ostrovski, John Quan, Remi Munos, Will Dabney

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 DM Lab-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

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).

\*\*\*\*\*

## Variational Sparse Coding

Francesco Tonolini, Bjorn Sand Jensen, Roderick Murray-Smith

### Variational auto-encoders

(VAEs) offer a tractable approach when performing approximate inference in otherwise intractable generative models. However, standard VAEs often produce latent codes that are disperse and lack interpretability, thus making the resulting representations unsuitable for auxiliary tasks (e.g. classification) and human inte

interpretation. We address these issues by merging ideas from variational auto-encoders and sparse coding, and propose to explicitly model sparsity in the latent space of a VAE with a Spike and Slab prior distribution. We derive the evidence lower bound using a discrete mixture recognition function thereby making approximate posterior inference as computationally efficient as in the standard VAE case. With the new approach, we are able to infer truly sparse representations with generally intractable non-linear probabilistic models. We show that these sparse representations are advantageous over standard VAE representations on two benchmark classification tasks (MNIST and Fashion-MNIST) by demonstrating improved classification accuracy and significantly increased robustness to the number of latent dimensions. Furthermore, we demonstrate qualitatively that the sparse elements capture subjectively understandable sources of variation.

\*\*\*\*\*

#### Canonical Correlation Analysis with Implicit Distributions

Yaxin Shi, Donna Xu, Yuangang Pan, Ivor Tsang

Canonical Correlation Analysis (CCA) is a ubiquitous technique that shows promising performance in multi-view learning problems. Due to the conjugacy of the prior and the likelihood, probabilistic CCA (PCCA) presents the posterior with an analytic solution, which provides probabilistic interpretation for classic linear CCA. As the multi-view data are usually complex in practice, nonlinear mappings are adopted to capture nonlinear dependency among the views. However, the interpretation provided in PCCA cannot be generalized to this nonlinear setting, as the distribution assumptions on the prior and the likelihood makes it restrictive to capture nonlinear dependency. To overcome this bottleneck, in this paper, we provide a novel perspective for CCA based on implicit distributions. Specifically, we present minimum Conditional Mutual Information (CMI) as a new criteria to capture nonlinear dependency for multi-view learning problem. To eliminate the explicit distribution requirement in direct estimation of CMI, we derive an objective whose minimization implicitly leads to the proposed criteria. Based on this objective, we present an implicit probabilistic formulation for CCA, named Implicit CCA (ICCA), which provides a flexible framework to design CCA extensions with implicit distributions. As an instantiation, we present adversarial CCA (ACCA), a nonlinear CCA variant which benefits from consistent encoding achieved by adversarial learning. Quantitative correlation analysis and superior performance on cross-view generation task demonstrate the superiority of the proposed ACCA.

\*\*\*\*\*

#### Multi-class classification without multi-class labels

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, Zsolt Kira

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.

\*\*\*\*\*

#### GraphSeq2Seq: Graph-Sequence-to-Sequence for Neural Machine Translation

Guoshuai Zhao, Jun Li, Lu Wang, Xueming Qian, Yun Fu

Sequence-to-Sequence (Seq2Seq) neural models have become popular for text generation problems, e.g. neural machine translation (NMT) (Bahdanau et al., 2014; Britz et al., 2017), text summarization (Nallapati et al., 2017; Wang & Ling, 2016), and image captioning (Venugopalan et al., 2015; Liu et al., 2017). Though sequential modeling has been shown to be effective, the dependency graph among words contains additional semantic information and thus can be utilized for sentence modeling. In this paper, we propose a Graph-Sequence-to-Seq

uence(GraphSeq2Seq) model to fuse the dependency graph among words into the traditional Seq2Seq framework. For each sample, the sub-graph of each word is encoded to a graph representation, which is then utilized to sequential encoding. At last, a sequence decoder is leveraged for output generation. Since above model fuses different features by contacting them together to encode, we also propose a variant of our model that regards the graph representations as additional annotations in attention mechanism (Bahdanau et al., 2014) by separately encoding different features. Experiments on several translation benchmarks show that our models can outperform existing state-of-the-art methods, demonstrating the effectiveness of the combination of Graph2Seq and Seq2Seq.

\*\*\*\*\*

Question Generation using a Scratchpad Encoder

Ryan Y Benmalek,Madian Khabisa,Suma Desu,Claire Cardie,Michele Banko

In this paper we introduce the Scratchpad Encoder, a novel addition to the sequence to sequence (seq2seq) framework and explore its effectiveness in generating natural language questions from a given logical form. The Scratchpad encoder enables the decoder at each time step to modify all the encoder outputs, thus using the encoder as a "scratchpad" memory to keep track of what has been generated so far and to guide future generation. Experiments on a knowledge based question generation dataset show that our approach generates more fluent and expressive questions according to quantitative metrics and human judgments.

\*\*\*\*\*

On the Spectral Bias of Neural Networks

Nasim Rahaman,Aristide Baratin,Devansh Arpit,Felix Draxler,Min Lin,Fred Hamprecht,Yoshua Bengio,Aaron Courville

Neural networks are known to be a class of highly expressive functions able to fit even random input-output mappings with 100% accuracy. In this work we present properties of neural networks that complement this aspect of expressivity. By using tools from Fourier analysis, we show that deep ReLU networks are biased towards low frequency functions, meaning that they cannot have local fluctuations without affecting their global behavior. Intuitively, this property is in line with the observation that over-parameterized networks find simple patterns that generalize across data samples. We also investigate how the shape of the data manifold affects expressivity by showing evidence that learning high frequencies gets easier with increasing manifold complexity, and present a theoretical understanding of this behavior. Finally, we study the robustness of the frequency components with respect to parameter perturbation, to develop the intuition that the parameters must be finely tuned to express high frequency functions.

\*\*\*\*\*

NEURAL MALWARE CONTROL WITH DEEP REINFORCEMENT LEARNING

Yu Wang,Jack W. Stokes,Mady Marinescu

Antimalware products are a key component in detecting malware attacks, and their engines typically execute unknown programs in a sandbox prior to running them on the native operating system. Files cannot be scanned indefinitely so the engine employs heuristics to determine when to halt execution. Previous research has investigated analyzing the sequence of system calls generated during this emulation process to predict if an unknown file is malicious, but these models require the emulation to be stopped after executing a fixed number of events from the beginning of the file. Also, these classifiers are not accurate enough to halt emulation in the middle of the file on their own. In this paper, we propose a novel algorithm which overcomes this limitation and learns the best time to halt the file's execution based on deep reinforcement learning (DRL). Because the new DRL-based system continues to emulate the unknown file until it can make a confident decision to stop, it prevents attackers from avoiding detection by initiating malicious activity after a fixed number of system calls. Results show that the proposed malware execution control model automatically halts emulation for 91.3% of the files earlier than heuristics employed by the engine. Furthermore, classifying the files at that time improves the true positive rate by 61.5%, at a false positive rate of 1%, compared to a baseline classifier.

\*\*\*\*\*

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

Sang-Woo Lee, Tong Gao, Sohee Yang, Jaejun Yoo, Jung-Woo Ha

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

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

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.

\*\*\*\*\*

## Area Attention

Yang Li, Lukasz Kaiser, Samy Bengio, Si Si

Existing attention mechanisms, are mostly item-based in that a model is trained to attend to individual items in a collection (the memory) where each item has a predefined, fixed granularity, e.g., a character or a word. Intuitively, an area in the memory consisting of multiple items can be worth attending to as a whole. We propose area attention: a way to attend to an area of the memory, where each area contains a group of items that are either spatially adjacent when the memory has a 2-dimensional structure, such as images, or temporally adjacent for 1-dimensional memory, such as natural language sentences. Importantly, the size of an area, i.e., the number of items in an area or the level of aggregation, is dynamically determined via learning, which can vary depending on the learned context.

erence of the adjacent items. By giving the model the option to attend to an area of items, instead of only individual items, a model can attend to information with varying granularity. Area attention can work along multi-head attention for attending to multiple areas in the memory. We evaluate area attention on two tasks: neural machine translation (both character and token-level) and image captioning, and improve upon strong (state-of-the-art) baselines in all the cases. These improvements are obtainable with a basic form of area attention that is parameter free. In addition to proposing the novel concept of area attention, we contribute an efficient way for computing it by leveraging the technique of summed area tables.

\*\*\*\*\*

Solving the Rubik's Cube with Approximate Policy Iteration

Stephen McAleer, Forest Agostinelli, Alexander Shmakov, Pierre Baldi

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.

\*\*\*\*\*

ACIQ: Analytical Clipping for Integer Quantization of neural networks

Ron Banner, Yury Nahshan, Elad Hoffer, Daniel Soudry

We analyze the trade-off between quantization noise and clipping distortion in low precision networks. We identify the statistics of various tensors, and derive exact expressions for the mean-square-error degradation due to clipping. By optimizing these expressions, we show marked improvements over standard quantization schemes that normally avoid clipping. For example, just by choosing the accurate clipping values, more than 40% accuracy improvement is obtained for the quantization of VGG-16 to 4-bits of precision. Our results have many applications for the quantization of neural networks at both training and inference time.

\*\*\*\*\*

Dynamic Channel Pruning: Feature Boosting and Suppression

Xitong Gao, Yiren Zhao, Lukasz Dudziak, Robert Mullins, Cheng-zhong Xu

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 5x and 2x 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  
 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 underlie 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

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 with Reflective Likelihoods

Adji B. Dieng, Kyunghyun Cho, David M. Blei, Yann LeCun

Models parameterized by deep neural networks have achieved state-of-the-art results in many domains. These models are usually trained using the maximum likelihood principle with a finite set of observations. However, training deep probabilistic models with maximum likelihood can lead to the issue we refer to as input forgetting. In deep generative latent-variable models, input forgetting corresponds to posterior collapse---a phenomenon in which the latent variables are driven independent from the observations. However input forgetting can happen even in the absence of latent variables. We attribute input forgetting in deep probabilistic models to the finite sample dilemma of maximum likelihood. We formalize this problem and propose a learning criterion---termed reflective likelihood---that explicitly prevents input forgetting. We empirically observe that the proposed criterion significantly outperforms the maximum likelihood objective when used in classification under a skewed class distribution. Furthermore, the reflective likelihood objective prevents posterior collapse when used to train stochastic auto-encoders with amortized inference. For example in a neural topic modeling experiment, the reflective likelihood objective leads to better qua



ntitative and qualitative results than the variational auto-encoder and the importance-weighted auto-encoder.

\*\*\*\*\*

On the effect of the activation function on the distribution of hidden nodes in a deep network

Philip M. Long and Hanie Sedghi

We analyze the joint probability distribution on the lengths of the vectors of hidden variables in different layers of a fully connected deep network, when the weights and biases are chosen randomly according to Gaussian distributions, and the input is binary-valued. We show that, if the activation function satisfies a minimal set of assumptions, satisfied by all activation functions that we know that are used in practice, then, as the width of the network gets large, the ``length process'' converges in probability to a length map that is determined as a simple function of the variances of the random weights and biases, and the activation function.

We also show that this convergence may fail for activation functions that violate our assumptions.

\*\*\*\*\*

Learning-Based Frequency Estimation Algorithms

Chen-Yu Hsu, Piotr Indyk, Dina Katabi, Ali Vakilian

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.

\*\*\*\*\*

Morpho-MNIST: Quantitative Assessment and Diagnostics for Representation Learning

Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, Ben Glocker

Revealing latent structure in data is an active field of research, having introduced exciting technologies such as variational autoencoders and adversarial networks, and is essential to push machine learning towards unsupervised knowledge discovery. However, a major challenge is the lack of suitable benchmarks for an objective and quantitative evaluation of learned representations. To address this issue we introduce Morpho-MNIST, a framework that aims to answer: "to what extent has my model learned to represent specific factors of variation in the data?"

We extend the popular MNIST dataset by adding a morphometric analysis enabling quantitative comparison of trained models, identification of the roles of latent variables, and characterisation of sample diversity. We further propose a set of quantifiable perturbations to assess the performance of unsupervised and supervised methods on challenging tasks such as outlier detection and domain adaptation.

\*\*\*\*\*

Asynchronous SGD without gradient delay for efficient distributed training

Roman Talyansky, Pavel Kisilev, Zach Melamed, Natan Peterfreund, Uri Verner

Asynchronous distributed gradient descent algorithms for training of deep neural networks are usually considered as inefficient, mainly because of the Gradient delay

problem. In this paper, we propose a novel asynchronous distributed algorithm that tackles this limitation by well-thought-out averaging of model updates, computed

by workers. The algorithm allows computing gradients along the process

of gradient merge, thus, reducing or even completely eliminating worker idle time due to communication overhead, which is a pitfall of existing asynchronous methods.

We provide theoretical analysis of the proposed asynchronous algorithm, and show its regret bounds. According to our analysis, the crucial parameter for keeping high convergence rate is the maximal discrepancy between local parameter vectors of any pair of workers. As long as it is kept relatively small, the convergence rate of the algorithm is shown to be the same as the one of a sequential

online learning. Furthermore, in our algorithm, this discrepancy is bounded by an expression that involves the staleness parameter of the algorithm, and is independent on the number of workers. This is the main differentiator between our approach and other solutions, such as Elastic Asynchronous SGD or Downpour SGD, in which that maximal discrepancy is bounded by an expression that depends on the number of workers, due to gradient delay problem. To demonstrate effectiveness of our approach, we conduct a series of experiments on image classification task on a cluster with 4 machines, equipped with a commodity communication

switch and with a single GPU card per machine. Our experiments show a linear scaling on 4-machine cluster without sacrificing the test accuracy

, while eliminating almost completely worker idle time. Since our method allows using commodity communication switch, it paves a way for large scale distributed training performed on commodity clusters.

\*\*\*\*\*

Transfer Learning for Related Reinforcement Learning Tasks via Image-to-Image Translation

Shani Gamrian, Yoav Goldberg

Deep Reinforcement Learning has managed to achieve state-of-the-art results in learning control policies directly from raw pixels. However, despite its remarkable success, it fails to generalize, a fundamental component required in a stable Artificial Intelligence system. Using the Atari game Breakout, we demonstrate the difficulty of a trained agent in adjusting to simple modifications in the raw image, ones that a human could adapt to trivially. In transfer learning, the goal is to use the knowledge gained from the source task to make the training of the target task faster and better. We show that using various forms of fine-tuning, a common method for transfer learning, is not effective for adapting to such small visual changes. In fact, it is often easier to re-train the agent from scratch than to fine-tune a trained agent. We suggest that in some cases transfer learning can be improved by adding a dedicated component whose goal is to learn to visually map between the known domain and the new one. Concretely, we use Unaligned Generative Adversarial Networks (GANs) to create a mapping function to translate images in the target task to corresponding images in the source task. These mapping functions allow us to transform between various variations of the Breakout game, as well as between different levels of a Nintendo game, Road Fighter. We show that learning this mapping is substantially more efficient than re-training. A visualization of a trained agent playing Breakout and Road Fighter, with and without the GAN transfer, can be seen in [url{https://streamable.com/msgtm}](https://streamable.com/msgtm) and [url{https://streamable.com/5e2ka}](https://streamable.com/5e2ka).

\*\*\*\*\*

BLISS in Non-Isometric Embedding Spaces

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, Graham Neubig  
Recent work on bilingual lexicon induction (BLI) has frequently depended either on aligned bilingual lexicons or on distribution matching, often with an assumption about the isometry of the two spaces. We propose a technique to quantitatively estimate this assumption of the isometry between two embedding spaces and empirically show that this assumption weakens as the languages in question become increasingly etymologically distant. We then propose Bilingual Lexicon Induction with Semi-Supervision (BLISS) --- a novel semi-supervised approach that relaxes

the isometric assumption while leveraging both limited aligned bilingual lexicons and a larger set of unaligned word embeddings, as well as a novel hubness filtering technique. Our proposed method improves over strong baselines for 11 of 14 on the MUSE dataset, particularly for languages whose embedding spaces do not appear to be isometric. In addition, we also show that adding supervision stabilizes the learning procedure, and is effective even with minimal supervision.

\*\*\*\*\*

Learning with Random Learning Rates.

Léonard Blier, Pierre Wolinski, Yann Ollivier

Hyperparameter tuning is a bothersome step in the training of deep learning models. One of the most sensitive hyperparameters is the learning rate of the gradient descent. We present the All Learning Rates At Once (Alrao) optimization method for neural networks: each unit or feature in the network gets its own learning rate sampled from a random distribution spanning several orders of magnitude.

This comes at practically no computational cost. Perhaps surprisingly, stochastic gradient descent (SGD) with Alrao performs close to SGD with an optimally tuned learning rate, for various architectures and problems. Alrao could save time when testing deep learning models: a range of models could be quickly assessed with Alrao, and the most promising models could then be trained more extensively. This text comes with a PyTorch implementation of the method, which can be plugged on an existing PyTorch model.

\*\*\*\*\*

Deep Ensemble Bayesian Active Learning : Addressing the Mode Collapse issue in Monte Carlo dropout via Ensembles

Remus Pop, Patric Fulop

In image classification tasks, the ability of deep convolutional neural networks (CNNs) to deal with complex image data has proved to be unrivalled. Deep CNNs, however, require large amounts of labeled training data to reach their full potential. In specialised domains such as healthcare, labeled data can be difficult and expensive to obtain. One way to alleviate this problem is to rely on active learning, a learning technique that aims to reduce the amount of labelled data needed for a specific task while still delivering satisfactory performance.

We propose a new active learning strategy designed

for deep neural networks. This method improves upon the current state-of-the-art deep Bayesian active learning method, which suffers from the mode collapse problem. We correct for this deficiency by making use of the expressive power and statistical properties of model ensembles. Our proposed method manages to capture superior data uncertainty, which translates into improved classification performance. We demonstrate empirically that our ensemble method yields faster convergence of CNNs trained on the MNIST and CIFAR-10 datasets.

\*\*\*\*\*

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

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 nonli

near 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.

\*\*\*\*\*

Look Ma, No GANs! Image Transformation with ModifAE

Chad Atalla, Bartholomew Tam, Amanda Song, Gary Cottrell

Existing methods of image to image translation require multiple steps in the training or modification process, and suffer from either an inability to generalize, or long training times. These methods also focus on binary trait modification, ignoring continuous traits. To address these problems, we propose ModifAE: a novel standalone neural network, trained exclusively on an autoencoding task, that implicitly learns to make continuous trait image modifications. As a standalone image modification network, ModifAE requires fewer parameters and less time to train than existing models. We empirically show that ModifAE produces significantly more convincing and more consistent continuous face trait modifications than the previous state-of-the-art model.

\*\*\*\*\*

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

Randall Balestriero, Richard Baraniuk

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 role 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  $\{\text{max-affine spline operators}\}$  (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.

$\{\text{This paper extends the MASO framework to these and an infinitely large class of new nonlinearities by linking deterministic MASOs with probabilistic Gaussian Mixture Models (GMMs).}\}$

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  $\text{``hard''}$  VQ inference problems, while sigmoid, hyperbolic tangent, and softmax can be interpreted as solutions to corresponding  $\text{``soft''}$  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  $\{\text{swish}\}$  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

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

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 Póczos, Aarti Singh

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.

\*\*\*\*\*

Domain Adaptation for Structured Output via Disentangled Patch Representations

Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, Manmohan Chandraker

Predicting structured outputs such as semantic segmentation relies on expensive per-pixel annotations to learn strong supervised models like convolutional neural networks. However, these models trained on one data domain may not generalize well to other domains unequipped with annotations for model finetuning. To avoid the labor-intensive process of annotation, we develop a domain adaptation method to adapt the source data to the unlabeled target domain. To this end, we propose to learn discriminative feature representations of patches based on label histograms in the source domain, through the construction of a disentangled space. With such representations as guidance, we then use an adversarial learning scheme to push the feature representations in target patches to the closer distributions in source ones. In addition, we show that our framework can integrate a global alignment process with the proposed patch-level alignment and achieve state-of-the-art performance on semantic segmentation. Extensive ablation studies and experiments are conducted on numerous benchmark datasets with various settings, such as synthetic-to-real and cross-city scenarios.

\*\*\*\*\*

Using Word Embeddings to Explore the Learned Representations of Convolutional Neural Networks

Dhanush Dharmaretnam,Chris Foster,Alona Fyshe

As deep neural net architectures minimize loss, they build up information in a hierarchy of learned representations that ultimately serve their final goal. Different architectures tackle this problem in slightly different ways, but all models aim to create representational spaces that accumulate information through the depth of the network. Here we build on previous work that indicated that two very different model classes trained on two very different tasks actually build knowledge representations that have similar underlying representations. Namely, we compare word embeddings from SkipGram (trained to predict co-occurring words) to several CNN architectures (trained for image classification) in order to understand how this accumulation of knowledge behaves in CNNs. We improve upon previous work by including 5 times more ImageNet classes in our experiments, and further expand the scope of the analyses to include a network trained on CIFAR-100. We characterize network behavior in pretrained models, and also during training, misclassification, and adversarial attack. Our work illustrates the power of using one model to explore another, gives new insights for CNN models, and provides a framework for others to perform similar analyses when developing new architectures.

\*\*\*\*\*

Implicit Maximum Likelihood Estimation

Ke Li,Jitendra Malik

Implicit probabilistic models are models defined naturally in terms of a sampling procedure and often induces a likelihood function that cannot be expressed explicitly. We develop a simple method for estimating parameters in implicit models that does not require knowledge of the form of the likelihood function or any derived quantities, but can be shown to be equivalent to maximizing likelihood under some conditions. Our result holds in the non-asymptotic parametric setting, where both the capacity of the model and the number of data examples are finite. We also demonstrate encouraging experimental results.

\*\*\*\*\*

Dopamine: A Research Framework for Deep Reinforcement Learning

Pablo Samuel Castro,Subhodeep Moitra,Carles Gelada,Saurabh Kumar,Marc G. Bellemare

Deep reinforcement learning (deep RL) research has grown significantly in recent years. A number of software offerings now exist that provide stable, comprehensive implementations for benchmarking. At the same time, recent deep RL research has become more diverse in its goals. In this paper we introduce Dopamine, a new research framework for deep RL that aims to support some of that diversity. Dopamine is open-source, TensorFlow-based, and provides compact yet reliable implementations of some state-of-the-art deep RL agents. We complement this offering with a taxonomy of the different research objectives in deep RL research. While by no means exhaustive, our analysis highlights the heterogeneity of research

ch

in the field, and the value of frameworks such as ours.

\*\*\*\*\*

#### A NOVEL VARIATIONAL FAMILY FOR HIDDEN NON-LINEAR MARKOV MODELS

Daniel Hernandez Diaz, Antonio Khalil Moretti, Ziqiang Wei, Shreya Saxena, John Cunningham, Liam Paninski

Latent variable models have been widely applied for the analysis and visualization of large datasets. In the case of sequential data, closed-form inference is possible when the transition and observation functions are linear. However, approximate inference techniques are usually necessary when dealing with nonlinear evolution and observations. Here, we propose a novel variational inference framework for the explicit modeling of time series, Variational Inference for Nonlinear Dynamics (VIND), that is able to uncover nonlinear observation and latent dynamics from sequential data. The framework includes a structured approximate posterior, and an algorithm that relies on the fixed-point iteration method to find the best estimate for latent trajectories. We apply the method to several datasets and show that it is able to accurately infer the underlying dynamics of these systems, in some cases substantially outperforming state-of-the-art methods.

\*\*\*\*\*

#### DppNet: Approximating Determinantal Point Processes with Deep Networks

Zelda Mariet, Jasper Snoek, Yaniv Ovadia

Determinantal Point Processes (DPPs) provide an elegant and versatile way to sample sets of items that balance the point-wise quality with the set-wise diversity of selected items. For this reason, they have gained prominence in many machine learning applications that rely on subset selection. However, sampling from a DPP over a ground set of size  $N$  is a costly operation, requiring in general an  $O(N^3)$  preprocessing cost and an  $O(Nk^3)$  sampling cost for subsets of size  $k$ . We approach this problem by introducing DppNets: generative deep models that produce DPP-like samples for arbitrary ground sets. We develop an inhibitive attention mechanism based on transformer networks that captures a notion of dissimilarity between feature vectors. We show theoretically that such an approximation is sensible as it maintains the guarantees of inhibition or dissimilarity that makes DPP so powerful and unique. Empirically, we demonstrate that samples from our model receive high likelihood under the more expensive DPP alternative.

\*\*\*\*\*

#### The Case for Full-Matrix Adaptive Regularization

Naman Agarwal, Brian Bullins, Xinyi Chen, Elad Hazan, Karan Singh, Cyril Zhang, Yi Zhang

Adaptive regularization methods pre-multiply a descent direction by a preconditioning matrix. Due to the large number of parameters of machine learning problems, full-matrix preconditioning methods are prohibitively expensive. We show how to modify full-matrix adaptive regularization in order to make it practical and effective. We also provide novel theoretical analysis for adaptive regularization in non-convex optimization settings. The core of our algorithm, termed GGT, consists of efficient inverse computation of square roots of low-rank matrices. Our preliminary experiments underscore improved convergence rate of GGT across a variety of synthetic tasks and standard deep learning benchmarks.

\*\*\*\*\*

#### Integral Pruning on Activations and Weights for Efficient Neural Networks

Qing Yang, Wei Wen, Zuoguan Wang, Yiran Chen, Hai Li

With the rapidly scaling up of deep neural networks (DNNs), extensive research studies on network model compression such as weight pruning have been performed for efficient deployment. This work aims to advance the compression beyond the weights to the activations of DNNs. We propose the Integral Pruning (IP) technique which integrates the activation pruning with the weight pruning. Through the learning on the different importance of neuron responses and connections, the generated network, namely IPnet, balances the sparsity between activations and weights and therefore further improves execution efficiency. The feasibility and effectiveness of IPnet are thoroughly evaluated through various network models with

different activation functions and on different datasets. With  $<0.5\%$  disturbance on the testing accuracy, IPnet saves  $71.1\% \sim 96.35\%$  of computation cost, compared to the original dense models with up to  $5.8x$  and  $10x$  reductions in activation and weight numbers, respectively.

\*\*\*\*\*

#### Bayesian Policy Optimization for Model Uncertainty

Gilwoo Lee, Brian Hou, Aditya Mandalika, Jeongseok Lee, Sanjiban Choudhury, Siddhartha S. Srinivasa

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.

\*\*\*\*\*

#### Towards Consistent Performance on Atari using Expert Demonstrations

Tobias Pohlen, Bilal Piot, Todd Hester, Mohammad Gheshlaghi Azar, Dan Horgan, David Budden, Gabriel Barth-Maron, Hado van Hasselt, John Quan, Mel Vecerík, Matteo Hessel, Rémi Munos, Olivier Pietquin

Despite significant advances in the field of deep Reinforcement Learning (RL), today's algorithms still fail to learn human-level policies consistently over a set of diverse tasks such as Atari 2600 games. We identify three key challenges that any algorithm needs to master in order to perform well on all games: processing diverse reward distributions, reasoning over long time horizons, and exploring efficiently. In this paper, we propose an algorithm that addresses each of these challenges and is able to learn human-level policies on nearly all Atari games. A new transformed Bellman operator allows our algorithm to process rewards of varying densities and scales; an auxiliary temporal consistency loss allows us to train stably using a discount factor of  $0.999$  (instead of  $0.99$ ) extending the effective planning horizon by an order of magnitude; and we ease the exploration problem by using human demonstrations that guide the agent towards rewarding states. When tested on a set of 42 Atari games, our algorithm exceeds the performance of an average human on 40 games using a common set of hyper parameters.

\*\*\*\*\*

#### Transformer-XL: Language Modeling with Longer-Term Dependency

Zihang Dai\*, Zhilin Yang\*, Yiming Yang, William W. Cohen, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov

We propose a novel neural architecture, Transformer-XL, for modeling longer-term dependency. To address the limitation of fixed-length contexts, we introduce a notion of recurrence by reusing the representations from the history. Empirically, we show state-of-the-art (SoTA) results on both word-level and character-level language modeling datasets, including WikiText-103, One Billion Word, Penn Treebank, and enwiki8. Notably, we improve the SoTA results from  $1.06$  to  $0.99$  in bpc on enwiki8, from  $33.0$  to  $18.9$  in perplexity on WikiText-103, and from  $28.0$  to  $23.5$  in perplexity on One Billion Word. Performance improves when the attention length increases during evaluation, and our best model attends to up to  $1,600$  words and  $3,800$  characters. To quantify the effective length of dependency, we devise a new metric and show that on WikiText-103 Transformer-XL manages to model dependency that is about  $80\%$  longer than recurrent networks and  $450\%$  longer than Transformer. Moreover, Transformer-XL is up to  $1,800+$  times faster than vanilla Transformer during evaluation.

\*\*\*\*\*



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

Sachin Kumar, Yulia Tsvetkov

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

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.

\*\*\*\*\*

ADAPTIVE NETWORK SPARSIFICATION VIA DEPENDENT VARIATIONAL BETA-BERNOULLI DROPOUT

Juho Lee, Saehoon Kim, Jaehong Yoon, Hae Beom Lee, Eunho Yang, Sung Ju Hwang

While variational dropout approaches have been shown to be effective for network sparsification, they are still suboptimal in the sense that they set the dropout rate for each neuron without consideration of the input data. With such input-independent dropout, each neuron is evolved to be generic across inputs, which makes it difficult to sparsify networks without accuracy loss. To overcome this limitation, we propose adaptive variational dropout whose probabilities are drawn from sparsity-inducing beta-Bernoulli prior. It allows each neuron to be evolved either to be generic or specific for certain inputs, or dropped altogether. Such input-adaptive sparsity-inducing dropout allows the resulting network to tolerate larger degree of sparsity without losing its expressive power by removing redundancies among features. We validate our dependent variational beta-Bernoulli dropout on multiple public datasets, on which it obtains significantly more compact networks than baseline methods, with consistent accuracy improvements over the base networks.

\*\*\*\*\*

Few-Shot Intent Inference via Meta-Inverse Reinforcement Learning

Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, Chelsea Finn

A significant challenge for the practical application of reinforcement learning to real world problems is the need to specify an oracle reward function that correctly defines a task. Inverse reinforcement learning (IRL) seeks to avoid this challenge by instead inferring a reward function from expert behavior. While appealing, it can be impractically expensive to collect datasets of demonstrations that cover the variation common in the real world (e.g. opening any type of door). Thus in practice, IRL must commonly be performed with only a limited set of d

emonstrations where it can be exceedingly difficult to unambiguously recover a reward function. In this work, we exploit the insight that demonstrations from other tasks can be used to constrain the set of possible reward functions by learning a "prior" that is specifically optimized for the ability to infer expressive reward functions from limited numbers of demonstrations. We demonstrate that our method can efficiently recover rewards from images for novel tasks and provide intuition as to how our approach is analogous to learning a prior.

\*\*\*\*\*

#### Physiological Signal Embeddings (PHASE) via Interpretable Stacked Models

Hugh Chen, Scott Lundberg, Gabe Erion, Su-In Lee

In health, machine learning is increasingly common, yet neural network embedding (representation) learning is arguably under-utilized for physiological signals.

This inadequacy stands out in stark contrast to more traditional computer science domains, such as computer vision (CV), and natural language processing (NLP). For physiological signals, learning feature embeddings is a natural solution to data insufficiency caused by patient privacy concerns -- rather than share data, researchers may share informative embedding models (i.e., representation models), which map patient data to an output embedding. Here, we present the PHASE (PHysiological Signal Embeddings) framework, which consists of three components: i) learning neural network embeddings of physiological signals, ii) predicting outcomes based on the learned embedding, and iii) interpreting the prediction results by estimating feature attributions in the "stacked" models (i.e., feature embedding model followed by prediction model). PHASE is novel in three ways: 1) To our knowledge, PHASE is the first instance of transferal of neural networks to create physiological signal embeddings. 2) We present a tractable method to obtain feature attributions through stacked models. We prove that our stacked model attributions can approximate Shapley values -- attributions known to have desirable properties -- for arbitrary sets of models. 3) PHASE was extensively tested in a cross-hospital setting including publicly available data. In our experiments, we show that PHASE significantly outperforms alternative embeddings -- such as raw, exponential moving average/variance, and autoencoder -- currently in use. Furthermore, we provide evidence that transferring neural network embedding/representation learners between distinct hospitals still yields performant embeddings and offer recommendations when transference is ineffective.

\*\*\*\*\*

#### DEEP ADVERSARIAL FORWARD MODEL

Morgan Funtowicz, Tomi Silander, Arnaud Sors, Julien Perez

Learning world dynamics has recently been investigated as a way to make reinforcement

learning (RL) algorithms to be more sample efficient and interpretable.

In this paper, we propose to capture an environment dynamics with a novel forward

model that leverages recent works on adversarial learning and visual control. Such

a model estimates future observations conditioned on the current ones and other input variables such as actions taken by an RL-agent. We focus on image generation

which is a particularly challenging topic but our method can be adapted to other modalities. More precisely, our forward model is trained to produce realistic

observations of the future while a discriminator model is trained to distinguish between real images and the model's prediction of the future. This approach overcomes

the need to define an explicit loss function for the forward model which is currently

used for solving such a class of problem. As a consequence, our learning protocol

does not have to rely on an explicit distance such as Euclidean distance which tends to produce unsatisfactory predictions. To illustrate our method, empirical qualitative and quantitative results are presented on a real driving scenario, a

long

with qualitative results on Atari game Frostbite.

\*\*\*\*\*

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

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.

\*\*\*\*\*

An Analysis of Composite Neural Network Performance from Function Composition Perspective

Ming-Chuan Yang, Meng Chang Chen

This work investigates the performance of a composite neural network, which is composed of pre-trained neural network models and non-instantiated neural network models, connected to form a rooted directed graph. A pre-trained neural network model is generally a well trained neural network model targeted for a specific function. The advantages of adopting such a pre-trained model in a composite neural network are two folds. One is to benefit from other's intelligence and diligence and the other is saving the efforts in data preparation and resources and time in training. However, the overall performance of composite neural network is still not clear. In this work, we prove that a composite neural network, with high probability, performs better than any of its pre-trained components under certain assumptions. In addition, if an extra pre-trained component is added to a composite network, with high probability the overall performance will be improved. In the empirical evaluations, distinctively different applications support the above findings.

\*\*\*\*\*

On Learning Heteroscedastic Noise Models within Differentiable Bayes Filters

Alina Kloss, Jeannette Bohg

In many robotic applications, it is crucial to maintain a belief about the state of

a system, like the location of a robot or the pose of an object.

These state estimates serve as input for planning and decision making and provide feedback during task execution.

Recursive Bayesian Filtering algorithms address the state estimation problem, but they require a model of the process dynamics and the sensory observations as well as

noise estimates that quantify the accuracy of these models.

Recently, multiple works have demonstrated that the process and sensor models can be

learned by end-to-end training through differentiable versions of Recursive Filtering methods.

However, even if the predictive models are known, finding suitable noise models remains challenging. Therefore, many practical applications rely on very simplistic noise

models.

Our hypothesis is that end-to-end training through differentiable Bayesian Filters enables us to learn more complex heteroscedastic noise models for the system dynamics. We evaluate learning such models with different types of filtering algorithms and on two different robotic tasks. Our experiments show that especially

for sampling-based filters like the Particle Filter, learning heteroscedastic noise models can drastically improve the tracking performance in comparison to using constant noise models.

\*\*\*\*\*

#### Exploring Curvature Noise in Large-Batch Stochastic Optimization

Yeming Wen, Kevin Luk, Maxime Gazeau, Guodong Zhang, Harris Chan, Jimmy Ba

Using stochastic gradient descent (SGD) with large batch-sizes to train deep neural networks is an increasingly popular technique. By doing so, one can improve parallelization by scaling to multiple workers (GPUs) and hence leading to significant reductions in training time. Unfortunately, a major drawback is the so-called generalization gap: large-batch training typically leads to a degradation in generalization performance of the model as compared to small-batch training. In this paper, we propose to correct this generalization gap by adding diagonal Fisher curvature noise to large-batch gradient updates. We provide a theoretical analysis of our method in the convex quadratic setting. Our empirical study with state-of-the-art deep learning models shows that our method not only improves the generalization performance in large-batch training but furthermore, does so in a way where the training convergence remains desirable and the training duration is not elongated. We additionally connect our method to recent works on loss surface landscape in the experimental section.

\*\*\*\*\*

#### From Nodes to Networks: Evolving Recurrent Neural Networks

Aditya Rawal, Jason Liang, Risto Miikkulainen

Gated recurrent networks such as those composed of Long Short-Term Memory (LSTM) nodes have recently been used to improve state of the art in many sequential

processing tasks such as speech recognition and machine translation. However, the basic structure of the LSTM node is essentially the same as when it was first conceived 25 years ago. Recently, evolutionary and reinforcement learning mechanisms have been employed to create new variations of this structure. This paper proposes a new method, evolution of a tree-based encoding of the gated memory nodes, and shows that it makes it possible to explore new variations more effectively than other methods. The method discovers nodes with multiple recurrent

paths and multiple memory cells, which lead to significant improvement in the standard language modeling benchmark task. Remarkably, this node did not perform well in another task, music modeling, but it was possible to evolve a different node that did, demonstrating that the approach discovers customized structure for

each task. The paper also shows how the search process can be speeded up by training an LSTM network to estimate performance of candidate structures, and by encouraging exploration of novel solutions. Thus, evolutionary design of complex

neural network structures promises to improve performance of deep learning architectures beyond human ability to do so.

\*\*\*\*\*

#### Generalized Tensor Models for Recurrent Neural Networks

Valentin Khrulkov, Oleksii Hrinchuk, Ivan Oseledets

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.

\*\*\*\*\*

#### Overcoming Multi-model Forgetting

Yassine Benyahia\*,Kaicheng Yu\*,Kamil Bennani-Smires,Martin Jaggi,Anthony Davison,  
Mathieu Salzmann,Claudiu Musat

We identify a phenomenon, which we refer to as *\*multi-model forgetting\**, that occurs when sequentially training multiple deep networks with partially-shared parameters; the performance of previously-trained models degrades as one optimizes a subsequent one, due to the overwriting of shared parameters. To overcome this, we introduce a statistically-justified weight plasticity loss that regularizes the learning of a model's shared parameters according to their importance for the previous models, and demonstrate its effectiveness when training two models sequentially and for neural architecture search. Adding weight plasticity in neural architecture search preserves the best models to the end of the search and yields improved results in both natural language processing and computer vision tasks.

\*\*\*\*\*

#### Sequenced-Replacement Sampling for Deep Learning

Chiu Man Ho,Dae Hoon Park,Wei Yang,Yi Chang

We propose sequenced-replacement sampling (SRS) for training deep neural networks. The basic idea is to assign a fixed sequence index to each sample in the dataset. Once a mini-batch is randomly drawn in each training iteration, we refill the original dataset by successively adding samples according to their sequence index. Thus we carry out replacement sampling but in a batched and sequenced way.

In a sense, SRS could be viewed as a way of performing "mini-batch augmentation". It is particularly useful for a task where we have a relatively small images-per-class such as CIFAR-100. Together with a longer period of initial large learning rate, it significantly improves the classification accuracy in CIFAR-100 over the current state-of-the-art results. Our experiments indicate that training deeper networks with SRS is less prone to over-fitting. In the best case, we achieve an error rate as low as 10.10%.

\*\*\*\*\*

#### NETWORK COMPRESSION USING CORRELATION ANALYSIS OF LAYER RESPONSES

Xavier Suau,Luca Zappella,Nicholas Apostoloff

Principal Filter Analysis (PFA) is an easy to implement, yet effective method for neural network compression. PFA exploits the intrinsic correlation between filter responses within network layers to recommend a smaller network footprint. We propose two compression algorithms: the first allows a user to specify the proportion of the original spectral energy that should be preserved in each layer after compression, while the second is a heuristic that leads to a parameter-free approach that automatically selects the compression used at each layer. Both algorithms are evaluated against several architectures and datasets, and we show considerable compression rates without compromising accuracy, e.g., for VGG-16 on CIFAR-10, CIFAR-100 and ImageNet, PFA achieves a compression rate of 8x, 3x, and 1.4x with an accuracy gain of 0.4%, 1.4% points, and 2.4% respectively. In our tests we also demonstrate that networks compressed with PFA achieve an accuracy that is very close to the empirical upper bound for a given compression ratio. Finally, we show how PFA is an effective tool for simultaneous compression and domain adaptation.

\*\*\*\*\*

#### Efficient Exploration through Bayesian Deep Q-Networks

Kamyar Azizzadenesheli,Animashree Anandkumar

We propose Bayesian Deep Q-Networks (BDQN), a principled and a practical Deep Reinforcement Learning (DRL) algorithm for Markov decision processes (MDP). It combines Thompson sampling with deep-Q networks (DQN). Thompson sampling ensures more efficient exploration-exploitation tradeoff in high dimensions. It is typically carried out through posterior sampling over the model parameters, which makes it computationally expensive. To overcome this limitation, we directly incorporate uncertainty over the value (Q) function. Further, we only introduce randomness in the last layer (i.e. the output layer) of the DQN and use independent Gaussian priors on the weights. This allows us to efficiently carry out Thompson sam

pling through Gaussian sampling and Bayesian Linear Regression (BLR), which has fast closed-form updates. The rest of the layers of the Q network are trained through back propagation, as in a standard DQN. We apply our method to a wide range of Atari games in Arcade Learning Environments and compare BDQN to a powerful baseline: the double deep Q-network (DDQN). Since BDQN carries out more efficient exploration, it is able to reach higher rewards substantially faster: in less than 5M+1M samples for almost half of the games to reach DDQN scores while a typical run of DDQN is 50-200M. We also establish theoretical guarantees for the special case when the feature representation is fixed and not learnt. We show that the Bayesian regret is bounded by  $O(\sqrt{d} \sqrt{N})$  after  $N$  time steps for a  $d$ -dimensional feature map, and this bound is shown to be tight up-to logarithmic factors. To the best of our knowledge, this is the first Bayesian theoretical guarantee for Markov Decision Processes (MDP) beyond the tabula rasa setting.

\*\*\*\*\*

#### Invariant and Equivariant Graph Networks

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, Yaron Lipman

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  $\text{linear}$  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  $2^k$  and  $15^k$ , respectively. More generally, for graph data defined on  $k$ -tuples of nodes, the dimension is the  $k$ -th and  $2^k$ -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

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.

\*\*\*\*\*

#### Combining Learned Representations for Combinatorial Optimization

Saavan Patel, Sayeef Salahuddin

We propose a new approach to combine Restricted Boltzmann Machines (RBMs) that can be used to solve combinatorial optimization problems. This allows synthesis of

f larger models from smaller RBMs that have been pretrained, thus effectively by passing the problem of learning in large RBMs, and creating a system able to model a large, complex multi-modal space. We validate this approach by using learned representations to create ``invertible boolean logic'', where we can use Markov chain Monte Carlo (MCMC) approaches to find the solution to large scale boolean satisfiability problems and show viability towards other combinatorial optimization problems. Using this method, we are able to solve 64 bit addition based problems, as well as factorize 16 bit numbers. We find that these combined representations can provide a more accurate result for the same sample size as compared to a fully trained model.

\*\*\*\*\*

EnGAN: Latent Space MCMC and Maximum Entropy Generators for Energy-based Models  
Rithesh Kumar, Anirudh Goyal, Aaron Courville, Yoshua Bengio

Unsupervised learning is about capturing dependencies between variables and is driven by the contrast between the probable vs improbable configurations of these variables, often either via a generative model which only samples probable ones or with an energy function (unnormalized log-density) which is low for probable ones and high for improbable ones. Here we consider learning both an energy function and an efficient approximate sampling mechanism for the corresponding distribution. Whereas the critic (or discriminator) in generative adversarial networks (GANs) learns to separate data and generator samples, introducing an entropy maximization regularizer on the generator can turn the interpretation of the critic into an energy function, which separates the training distribution from everything else, and thus can be used for tasks like anomaly or novelty detection.

This paper is motivated by the older idea of sampling in latent space rather than data space because running a Monte-Carlo Markov Chain (MCMC) in latent space has been found to be easier and more efficient, and because a GAN-like generator can convert latent space samples to data space samples. For this purpose, we show how a Markov chain can be run in latent space whose samples can be mapped to data space, producing better samples. These samples are also used for the negative phase gradient required to estimate the log-likelihood gradient of the data space energy function. To maximize entropy at the output of the generator, we take advantage of recently introduced neural estimators of mutual information. We find that in addition to producing a useful scoring function for anomaly detection, the resulting approach produces sharp samples (like GANs) while covering the modes well, leading to high Inception and Fréchet scores.

\*\*\*\*\*

Cohen Welling bases & SO(2)-Equivariant classifiers using Tensor nonlinearity.  
Muthuvel Murugan, K Venkata Subrahmanyam

In this paper we propose autoencoder architectures for learning a Cohen-Welling (CW)-basis for images and their rotations. We use the learned CW-basis to build a rotation equivariant classifier to classify images. The autoencoder and classi-

-fier architectures use only tensor product nonlinearity. The model proposed by Cohen & Welling (2014) uses ideas from group representation theory, and extracts a basis exposing irreducible representations for images and their rotations. We give several architectures to learn CW-bases including a novel coupling AE archi-

- tecture to learn a coupled CW-bases for images in different scales simultaneously.

Our use of tensor product nonlinearity is inspired from recent work of Kondor (2018a). Our classifier has very good accuracy and we use fewer parameters. Even when the sample complexity to learn a good CW-basis is low we learn classifiers which perform impressively. We show that a coupled CW-bases in one scale can be deployed to classify images in a classifier trained and tested on images in

a different scale with only a marginal dip in performance.

\*\*\*\*\*

## Residual Non-local Attention Networks for Image Restoration

Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, Yun Fu

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.

\*\*\*\*\*

## Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil

A significant roadblock in multilingual neural language modeling is the lack of labeled non-English data. One potential method for overcoming this issue is learning cross-lingual text representations that can be used to transfer the performance from training on English tasks to non-English tasks, despite little to no task-specific non-English data. In this paper, we explore a natural setup for learning crosslingual sentence representations: the dual-encoder. We provide a comprehensive evaluation of our cross-lingual representations on a number of monolingual, crosslingual, and zero-shot/few-shot learning tasks, and also give an analysis of different learned cross-lingual embedding spaces.

\*\*\*\*\*

## Kernel RNN Learning (KeRNL)

Christopher Roth, Ingmar Kanitscheider, Ila Fiete

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

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



either 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, Yanzi Wang, Xue Lin

When generating adversarial examples to attack deep neural networks (DNNs),  $L_p$  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  $L_p$ -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 (Papier et al., 2016b) and class activation map (Zhou et al., 2016).

\*\*\*\*\*

Dissecting an Adversarial framework for Information Retrieval

Ameet Deshpande, Mitesh M. Khapra

Recent advances in Generative Adversarial Networks facilitated by improvements to the framework and successful application to various problems has resulted in extensions to multiple domains. IRGAN attempts to leverage the framework for Information-Retrieval (IR), a task that can be described as modeling the conditional probability distribution  $p(d|q)$  over the documents ( $d$ ), given the query ( $q$ ). The work that proposes IRGAN claims that optimizing their minimax loss function will result in a generator which can learn the distribution, but their setup and baseline term steer the model away from an exact adversarial formulation, and this work attempts to point out certain inaccuracies in their formulation. Analyzing their loss curves gives insight into possible mistakes in the loss functions and better performance can be obtained by using the co-training like setup we propose, where two models are trained in a co-operative rather than an adversarial fashion.

\*\*\*\*\*

Complexity of Training ReLU Neural Networks

Digvijay Boob, Santanu S. Dey, Guanghui Lan

In this paper, we explore some basic questions on complexity of training Neural networks with ReLU activation function. We show that it is NP-hard to train a two-hidden layer feedforward ReLU neural network. If dimension  $d$  of the data is fixed then we show that there exists a polynomial time algorithm for the same training problem. We also show that if sufficient over-parameterization is provided in the first hidden layer of ReLU neural network then there is a polynomial time algorithm which finds weights such that output of the over-parameterized ReLU neural network matches with the output of the given data.

\*\*\*\*\*

Graph Convolutional Network with Sequential Attention For Goal-Oriented Dialogue Systems

Suman Banerjee, Mitesh M. Khapra

Domain specific goal-oriented dialogue systems typically require modeling three types of inputs, viz., (i) the knowledge-base associated with the domain, (ii) the history of the conversation, which is a sequence of utterances and (iii) the

current utterance for which the response needs to be generated. While modeling these inputs, current state-of-the-art models such as Mem2Seq typically ignore the rich structure inherent in the knowledge graph and the sentences in the conversation context. Inspired by the recent success of structure-aware Graph Convolutional Networks (GCNs) for various NLP tasks such as machine translation, semantic role labeling and document dating, we propose a memory augmented GCN for goal-oriented dialogues. Our model exploits (i) the entity relation graph in a knowledge-base and (ii) the dependency graph associated with an utterance to compute richer representations for words and entities. Further, we take cognizance of the fact that in certain situations, such as, when the conversation is in a code-mixed language, dependency parsers may not be available. We show that in such situations we could use the global word co-occurrence graph and use it to enrich the representations of utterances. We experiment with the modified DSTC2 dataset and its recently released code-mixed versions in four languages and show that our method outperforms existing state-of-the-art methods, using a wide range of evaluation metrics.

\*\*\*\*\*

#### Cross-Entropy Loss Leads To Poor Margins

Kamil Nar,Orhan Ocal,S. Shankar Sastry,Kannan Ramchandran

Neural networks could misclassify inputs that are slightly different from their training data, which indicates a small margin between their decision boundaries and the training dataset. In this work, we study the binary classification of linearly separable datasets and show that linear classifiers could also have decision boundaries that lie close to their training dataset if cross-entropy loss is used for training. In particular, we show that if the features of the training dataset lie in a low-dimensional affine subspace and the cross-entropy loss is minimized by using a gradient method, the margin between the training points and the decision boundary could be much smaller than the optimal value. This result is contrary to the conclusions of recent related works such as (Soudry et al., 2018), and we identify the reason for this contradiction. In order to improve the margin, we introduce differential training, which is a training paradigm that uses a loss function defined on pairs of points from each class. We show that the decision boundary of a linear classifier trained with differential training indeed achieves the maximum margin. The results reveal the use of cross-entropy loss as one of the hidden culprits of adversarial examples and introduces a new direction to make neural networks robust against them.

\*\*\*\*\*

#### Learning to Refer to 3D Objects with Natural Language

Panos Achlioptas,Judy E. Fan,Robert X.D. Hawkins,Noah D. Goodman,Leo Guibas

Human world knowledge is both structured and flexible. When people see an object, they represent it not as a pixel array but as a meaningful arrangement of semantic parts. Moreover, when people refer to an object, they provide descriptions that are not merely true but also relevant in the current context. Here, we combine these two observations in order to learn fine-grained correspondences between language and contextually relevant geometric properties of 3D objects. To do this, we employed an interactive communication task with human participants to construct a large dataset containing natural utterances referring to 3D objects from ShapeNet in a wide variety of contexts. Using this dataset, we developed neural listener and speaker models with strong capacity for generalization. By performing targeted lesions of visual and linguistic input, we discovered that the neural listener depends heavily on part-related words and associates these words correctly with the corresponding geometric properties of objects, suggesting that it has learned task-relevant structure linking the two input modalities. We further show that a neural speaker that is 'listener-aware' --- that plans its utterances according to how an imagined listener would interpret its words in context --- produces more discriminative referring expressions than an 'listener-unaware' speaker, as measured by human performance in identifying the correct object.

\*\*\*\*\*

#### Stochastic Adversarial Video Prediction

Alex X. Lee,Richard Zhang,Frederik Ebert,Pieter Abbeel,Chelsea Finn,Sergey Levin

e

Being able to predict what may happen in the future requires an in-depth understanding of the physical and causal rules that govern the world. A model that is able to do so has a number of appealing applications, from robotic planning to representation learning. However, learning to predict raw future observations, such as frames in a video, is exceedingly challenging—the ambiguous nature of the problem can cause a naively designed model to average together possible futures into a single, blurry prediction. Recently, this has been addressed by two distinct approaches: (a) latent variational variable models that explicitly model underlying stochasticity and (b) adversarially-trained models that aim to produce naturalistic images. However, a standard latent variable model can struggle to produce realistic results, and a standard adversarially-trained model underutilizes latent variables and fails to produce diverse predictions. We show that these distinct methods are in fact complementary. Combining the two produces predictions that look more realistic to human raters and better cover the range of possible futures. Our method outperforms prior works in these aspects.

\*\*\*\*\*

Beyond Winning and Losing: Modeling Human Motivations and Behaviors with Vector-valued Inverse Reinforcement Learning

Baoxiang Wang, Tongfang Sun, Xianjun Sam Zheng

In recent years, reinforcement learning methods have been applied to model gameplay with great success, achieving super-human performance in various environments, such as Atari, Go and Poker.

However, those studies mostly focus on winning the game and have largely ignored the rich and complex human motivations, which are essential for understanding the agents' diverse behavior.

In this paper, we present a multi-motivation behavior modeling which investigates the multifaceted human motivations and models the underlying value structure of the agents.

Our approach extends inverse RL to the vectorized-valued setting which imposes a much weaker assumption than previous studies.

The vectorized rewards incorporate Pareto optimality, which is a powerful tool to explain a wide range of behavior by its optimality.

For practical assessment, our algorithm is tested on the World of Warcraft Avatar History dataset spanning three years of the gameplay.

Our experiments demonstrate the improvement over the scalarization-based methods on real-world problem settings.

\*\*\*\*\*

Model Comparison for Semantic Grouping

Francisco Vargas, Kamen Brestnichki, Nils Hammerla

We introduce a probabilistic framework for quantifying the semantic similarity between two groups of embeddings. We formulate the task of semantic similarity as a model comparison task in which we contrast a generative model which jointly models two sentences versus one that does not. We illustrate how this framework can be used for the Semantic Textual Similarity tasks using clear assumptions about how the embeddings of words are generated. We apply information criteria based model comparison to overcome the shortcomings of Bayesian model comparison, whilst still penalising model complexity. We achieve competitive results by applying the proposed framework with an appropriate choice of likelihood on the STS datasets.

\*\*\*\*\*

High Resolution and Fast Face Completion via Progressively Attentive GANs

Zeyuan Chen, Shaoliang Nie, Tianfu Wu, Christopher G. Healey

Face completion is a challenging task with the difficulty level increasing significantly with respect to high resolution, the complexity of "holes" and the controllable attributes of filled-in fragments. Our system addresses the challenges by learning a fully end-to-end framework that trains generative adversarial networks (GANs) progressively from low resolution to high resolution with conditional vectors encoding controllable attributes. We design a novel coarse-to-fine attentive module network architecture. Our model is encouraged to attend on finer details.

etails while the network is growing to a higher resolution, thus being capable of showing progressive attention to different frequency components in a coarse-to-fine way. We term the module Frequency-oriented Attentive Module (FAM). Our system can complete faces with large structural and appearance variations using a single feed-forward pass of computation with mean inference time of 0.54 seconds for images at 1024x1024 resolution. A pilot human study shows our approach outperforms state-of-the-art face completion methods. The code will be released upon publication.

\*\*\*\*\*

#### Robust Determinantal Generative Classifier for Noisy Labels and Adversarial Attacks

Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, Jinwoo Shin

Large-scale datasets may contain significant proportions of noisy (incorrect) class labels, and it is well-known that modern deep neural networks poorly generalize from such noisy training datasets. In this paper, we propose a novel inference method, Deep Determinantal Generative Classifier (DDGC), which can obtain a more robust decision boundary under any softmax neural classifier pre-trained on noisy datasets. Our main idea is inducing a generative classifier on top of hidden feature spaces of the discriminative deep model. By estimating the parameters of generative classifier using the minimum covariance determinant estimator, we significantly improve the classification accuracy, with neither re-training of the deep model nor changing its architectures. In particular, we show that DDGC not only generalizes well from noisy labels, but also is robust against adversarial perturbations due to its large margin property. Finally, we propose the ensemble version of DDGC to improve its performance, by investigating the layer-wise characteristics of generative classifier. Our extensive experimental results demonstrate the superiority of DDGC given different learning models optimized by various training techniques to handle noisy labels or adversarial samples. For instance, on CIFAR-10 dataset containing 45% noisy training labels, we improve the test accuracy of a deep model optimized by the state-of-the-art noise-handling training method from 33.34% to 43.02%.

\*\*\*\*\*

#### Dense Morphological Network: An Universal Function Approximator

Ranjan Mondal, Sanchayan Santra, Bhabatosh Chanda

Artificial neural networks are built on the basic operation of linear combination and non-linear activation function. Theoretically this structure can approximate any continuous function with three layer architecture. But in practice learning the parameters of such network can be hard. Also the choice of activation function can greatly impact the performance of the network. In this paper we are proposing to replace the basic linear combination operation with non-linear operations that do away with the need of additional non-linear activation function. To this end we are proposing the use of elementary morphological operations (dilation and erosion) as the basic operation in neurons. We show that these networks (Denoted as Morph-Net) with morphological operations can approximate any smooth function requiring less number of parameters than what is necessary for normal neural networks. The results show that our network perform favorably when compared with similar structured network. We have carried out our experiments on MNIST, Fashion-MNIST, CIFAR10 and CIFAR100.

\*\*\*\*\*

#### Discriminative out-of-distribution detection for semantic segmentation

Petra Bevandi<sup>1</sup>, Siniša Šegvić<sup>2</sup>, Ivan Krešo, Marin Oršić<sup>3</sup>

Most classification and segmentation datasets assume a closed-world scenario in which predictions are expressed as distribution over a predetermined set of visual classes. However, such assumption implies unavoidable and often unnoticeable failures in presence of out-of-distribution (OOD) input. These failures are bound to happen in most real-life applications since current visual ontologies are far from being comprehensive. We propose to address this issue by discriminative detection

of OOD pixels in input data. Different from recent approaches, we avoid to bring any decisions by only observing the training dataset of the primary model train

ed to solve the desired computer vision task. Instead, we train a dedicated OOD model

which discriminates the primary training set from a much larger "background" dataset which approximates the variety of the visual world. We perform our experiments on high resolution natural images in a dense prediction setup. We use several road driving datasets as our training distribution, while we approximate the background distribution with the ILSVRC dataset. We evaluate our approach on WildDash test, which is currently the only public test dataset with out-of-distribution images.

The obtained results show that the proposed approach succeeds to identify out-of-distribution pixels while outperforming previous work by a wide margin.

\*\*\*\*\*

Direct Optimization through  $\arg \max$  for Discrete Variational Auto-Encoder  
Guy Lorberbom, Tamir Hazan

Reparameterization of variational auto-encoders is an effective method for reducing the variance of their gradient estimates. However, when the latent variables are discrete, a reparameterization is problematic due to discontinuities in the discrete space. In this work, we extend the direct loss minimization technique to discrete variational auto-encoders. We first reparameterize a discrete random variable using the  $\arg \max$  function of the Gumbel-Max perturbation model. We then use direct optimization to propagate gradients through the non-differentiable  $\arg \max$  using two perturbed  $\arg \max$  operations.

\*\*\*\*\*

Lorentzian Distance Learning

Marc T Law, Jake Snell, Richard S Zemel

This paper introduces an approach to learn representations based on the Lorentzian distance in hyperbolic geometry. Hyperbolic geometry is especially suited to hierarchically-structured datasets, which are prevalent in the real world. Current hyperbolic representation learning methods compare examples with the Poincaré distance metric. They formulate the problem as minimizing the distance of each node in a hierarchy with its descendants while maximizing its distance with other nodes. This formulation produces node representations close to the centroid of their descendants. We exploit the fact that the centroid w.r.t the squared Lorentzian distance can be written in closed-form. We show that the Euclidean norm of such a centroid decreases as the curvature of the hyperbolic space decreases. This property makes it appropriate to represent hierarchies where parent nodes minimize the distances to their descendants and have smaller Euclidean norm than their children. Our approach obtains state-of-the-art results in retrieval and classification tasks on different datasets.

\*\*\*\*\*

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

Alon Jacovi, Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, Jonathan Berant

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.

\*\*\*\*\*

#### Stackelberg GAN: Towards Provable Minimax Equilibrium via Multi-Generator Architectures

Hongyang Zhang, Susu Xu, Jiantao Jiao, Pengtao Xie, Ruslan Salakhutdinov, Eric P. Xing

We study the problem of alleviating the instability issue in the GAN training procedure via new architecture design. The discrepancy between the minimax and maximin objective values could serve as a proxy for the difficulties that the alternating gradient descent encounters in the optimization of GANs. In this work, we give new results on the benefits of multi-generator architecture of GANs. We show that the minimax gap shrinks to  $\epsilon$  as the number of generators increases with rate  $O(1/\epsilon)$ . This improves over the best-known result of  $O(1/\epsilon^2)$ . At the core of our techniques is a novel application of Shapley-Folkman lemma to the generic minimax problem, where in the literature the technique was only known to work when the objective function is restricted to the Lagrangian function of a constraint optimization problem. Our proposed Stackelberg GAN performs well experimentally in both synthetic and real-world datasets, improving Fréchet Inception Distance by 14.61% over the previous multi-generator GANs on the benchmark datasets.

\*\*\*\*\*

#### RNNs implicitly implement tensor-product representations

R. Thomas McCoy, Tal Linzen, Ewan Dunbar, Paul Smolensky

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 by TPRs; 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, Zolt Kira, Richard Socher, Caiming Xiong

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

ough 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>.

\*\*\*\*\*

#### Negotiating Team Formation Using Deep Reinforcement Learning

Yoram Bachrach, Richard Everett, Edward Hughes, Angeliki Lazaridou, Joel Leibo, Marc Lanctot, Mike Johanson, Wojtek Czarnecki, Thore Graepel

When autonomous agents interact in the same environment, they must often cooperate to achieve their goals. One way for agents to cooperate effectively is to form a team, make a binding agreement on a joint plan, and execute it. However, when agents are self-interested, the gains from team formation must be allocated appropriately to incentivize agreement. Various approaches for multi-agent negotiation have been proposed, but typically only work for particular negotiation protocols. More general methods usually require human input or domain-specific data, and so do not scale. To address this, we propose a framework for training agents to negotiate and form teams using deep reinforcement learning. Importantly, our method makes no assumptions about the specific negotiation protocol, and is instead completely experience driven. We evaluate our approach on both non-spatial and spatially extended team-formation negotiation environments, demonstrating that our agents beat hand-crafted bots and reach negotiation outcomes consistent with fair solutions predicted by cooperative game theory. Additionally, we investigate how the physical location of agents influences negotiation outcomes.

\*\*\*\*\*

#### Improved robustness to adversarial examples using Lipschitz regularization of the loss

Chris Finlay, Adam M. Oberman, Bilal Abbasi

We augment adversarial training (AT) with worst case adversarial training (WCAT) which improves adversarial robustness by 11% over the current state-of-the-art result in the  $\ell_2$ -norm on CIFAR-10. We interpret adversarial training as

Total Variation Regularization, which is a fundamental tool in mathematical image processing, and WCAT as Lipschitz regularization, which appears in Image Inpainting. We obtain verifiable worst and average case robustness guarantees, based on the expected and maximum values of the norm of the gradient of the loss.

\*\*\*\*\*

#### EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez Lobato, Sebastian Nowozin, Cheng Zhang

Making decisions requires information relevant to the task at hand. Many real-life decision-making situations allow acquiring further relevant information at a specific cost. For example, in assessing the health status of a patient we may decide to take additional measurements such as diagnostic tests or imaging scans before making a final assessment. More information that is relevant allows for better decisions but it may be costly to acquire all of this information. How can we trade off the desire to make good decisions with the option to acquire further information at a cost? To this end, we propose a principled framework, named EDDI (Efficient Dynamic Discovery of high-value Information), based on the theory of Bayesian experimental design. In EDDI we propose a novel partial variational autoencoder (Partial VAE), to efficiently handle missing data over varying subsets of known information. EDDI combines this Partial VAE with an acquisition function that maximizes expected information gain on a set of target variables. EDDI is efficient and demonstrates that dynamic discovery of high-value information is possible; we show cost reduction at the same decision quality and improved decision quality at the same cost in benchmarks and in two health-care applications. We believe there is great potential for realizing these gains in real-world decision support systems.

\*\*\*\*\*

#### Policy Transfer with Strategy Optimization

Wenhao Yu,C. Karen Liu,Greg Turk

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.

\*\*\*\*\*

Predicted Variables in Programming

Victor Carbune,Thierry Coppey,Alexander Daryin,Thomas Deselaers,Nikhil Sarda,Jay Yagnik

We present Predicted Variables, an approach to making machine learning (ML) a first class citizen in programming languages.

There is a growing divide in approaches to building systems: using human experts (e.g. programming) on the one hand, and using behavior learned from data (e.g. ML) on the other hand. PVars aim to make using ML in programming easier by hybridizing the two. We leverage the existing concept of variables and create a new type, a predicted variable. PVars are akin to native variables with one important distinction: PVars determine their value using ML when evaluated. We describe PVars and their interface, how they can be used in programming, and demonstrate the feasibility of our approach on three algorithmic problems: binary search, QuickSort, and caches.

We show experimentally that PVars are able to improve over the commonly used heuristics and lead to a better performance than the original algorithms.

As opposed to previous work applying ML to algorithmic problems, PVars have the advantage that they can be used within the existing frameworks and do not require the existing domain knowledge to be replaced. PVars allow for a seamless integration of ML into existing systems and algorithms.

Our PVars implementation currently relies on standard Reinforcement Learning (RL) methods. To learn faster, PVars use the heuristic function, which they are replacing, as an initial function. We show that PVars quickly pick up the behavior of the initial function and then improve performance beyond that without ever performing substantially worse -- allowing for a safe deployment in critical applications.

\*\*\*\*\*

DeepOBS: A Deep Learning Optimizer Benchmark Suite

Frank Schneider,Lukas Balles,Philipp Hennig

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 Dee



POBS, 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.

\*\*\*\*\*

#### EFFICIENT SEQUENCE LABELING WITH ACTOR-CRITIC TRAINING

Saeed Najafi, Colin Cherry, Greg Kondrak

Neural approaches to sequence labeling often use a Conditional Random Field (CRF) to model their output dependencies, while Recurrent Neural Networks (RNN) are used for the same purpose in other tasks. We set out to establish RNNs as an attractive alternative to CRFs for sequence labeling. To do so, we address one of the RNN's most prominent shortcomings, the fact that it is not exposed to its own errors with the maximum-likelihood training. We frame the prediction of the output sequence as a sequential decision-making process, where we train the network with an adjusted actor-critic algorithm (AC-RNN). We comprehensively compare this strategy with maximum-likelihood training for both RNNs and CRFs on three structured-output tasks. The proposed AC-RNN efficiently matches the performance of the CRF on NER and CCG tagging, and outperforms it on Machine Transliteration. We also show that our training strategy is significantly better than other techniques for addressing RNN's exposure bias, such as Scheduled Sampling, and Self-Critical policy training.

\*\*\*\*\*

#### Transferrable End-to-End Learning for Protein Interface Prediction

Raphael J. L. Townshend, Rishi Bedi, Ron O. Dror

While there has been an explosion in the number of experimentally determined, atomically detailed structures of proteins, how to represent these structures in a machine learning context remains an open research question. In this work we demonstrate that representations learned from raw atomic coordinates can outperform hand-engineered structural features while displaying a much higher degree of transferrability. To do so, we focus on a central problem in biology: predicting how proteins interact with one another—that is, which surfaces of one protein bind to which surfaces of another protein. We present Siamese Atomic Surfacelet Network (SASNet), the first end-to-end learning method for protein interface prediction. Despite using only spatial coordinates and identities of atoms as inputs, SASNet outperforms state-of-the-art methods that rely on hand-engineered, high-level features. These results are particularly striking because we train the method entirely on a significantly biased data set that does not account for the fact that proteins deform when binding to one another. Demonstrating the first successful application of transfer learning to atomic-level data, our network maintains high performance, without retraining, when tested on real cases in which proteins do deform.

\*\*\*\*\*

#### Learning data-derived privacy preserving representations from information metrics

Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Guillermo Sapiro

It is clear that users should own and control their data and privacy. Utility providers are also becoming more interested in guaranteeing data privacy. Therefore, users and providers can and should collaborate in privacy protecting challenges, and this paper addresses this new paradigm. We propose a framework where the user controls what characteristics of the data they want to share (utility) and what they want to keep private (secret), without necessarily asking the utility

provider to change its existing machine learning algorithms. We first analyze the space of privacy-preserving representations and derive natural information-theoretic bounds on the utility-privacy trade-off when disclosing a sanitized version of the data  $X$ . We present explicit learning architectures to learn privacy-preserving representations that approach this bound in a data-driven fashion. We describe important use-case scenarios where the utility providers are willing to collaborate with the sanitization process. We study space-preserving transformations where the utility provider can use the same algorithm on original and sanitized data, a critical and novel attribute to help service providers accommodate varying privacy requirements with a single set of utility algorithms. We illustrate this framework through the implementation of three use cases; subject-within-subject, where we tackle the problem of having a face identity detector that works only on a consenting subset of users, an important application, for example, for mobile devices activated by face recognition; gender-and-subject, where we preserve facial verification while hiding the gender attribute for users who choose to do so; and emotion-and-gender, where we hide independent variables, as is the case of hiding gender while preserving emotion detection.

\*\*\*\*\*

Graph Spectral Regularization For Neural Network Interpretability

Alexander Tong, David van Dijk, Jay Stanley, Guy Wolf, Smriti Krishnaswamy

Deep neural networks can learn meaningful representations of data. However, these representations are hard to interpret. For example, visualizing a latent layer is generally only possible for at most three dimensions. Neural networks are able to learn and benefit from much higher dimensional representations but these are not visually interpretable because nodes have arbitrary ordering within a layer. Here, we utilize the ability of the human observer to identify patterns in structured representations to visualize higher dimensions. To do so, we propose a class of regularizations we call \textit{Graph Spectral Regularizations} that impose graph-structure on latent layers. This is achieved by treating activations as signals on a predefined graph and constraining those activations using graph filters, such as low pass and wavelet-like filters. This framework allows for any kind of graph as well as filter to achieve a wide range of structured regularizations depending on the inference needs of the data. First, we show a synthetic example that the graph-structured layer can reveal topological features of the data. Next, we show that a smoothing regularization can impose semantically consistent ordering of nodes when applied to capsule nets. Further, we show that the graph-structured layer, using wavelet-like spatially localized filters, can form localized receptive fields for improved image and biomedical data interpretation. In other words, the mapping between latent layer, neurons and the output space becomes clear due to the localization of the activations. Finally, we show that when structured as a grid, the representations create coherent images that allow for image-processing techniques such as convolutions.

\*\*\*\*\*

signSGD with Majority Vote is Communication Efficient and Fault Tolerant

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, Anima Anandkumar

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.

\*\*\*\*\*

#### Task-GAN for Improved GAN based Image Restoration

Jiahong Ouyang, Guanhua Wang, Enhao Gong, Kevin Chen, John Pauly and Greg Zaharchuk  
Deep Learning (DL) algorithms based on Generative Adversarial Network (GAN) have demonstrated great potentials in computer vision tasks such as image restoration. Despite the rapid development of image restoration algorithms using DL and GANs, image restoration for specific scenarios, such as medical image enhancement and super-resolved identity recognition, are still facing challenges. How to ensure visually realistic restoration while avoiding hallucination or mode-collapse? How to make sure the visually plausible results do not contain hallucinated features jeopardizing downstream tasks such as pathology identification and subject identification?

Here we propose to resolve these challenges by coupling the GAN based image restoration framework with another task-specific network. With medical imaging restoration as an example, the proposed model conducts additional pathology recognition/classification task to ensure the preservation of detailed structures that are important to this task. Validated on multiple medical datasets, we demonstrate the proposed method leads to improved deep learning based image restoration while preserving the detailed structure and diagnostic features. Additionally, the trained task network show potentials to achieve super-human level performance in identifying pathology and diagnosis.

Further validation on super-resolved identity recognition tasks also show that the proposed method can be generalized for diverse image restoration tasks.

\*\*\*\*\*

#### Set Transformer

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosior, Seungjin Choi, Yee Whye Teh  
Many machine learning tasks such as multiple instance learning, 3D shape recognition and few-shot image classification are defined on sets of instances. Since solutions to such problems do not depend on the permutation of elements of the set, models used to address them should be permutation invariant. We present an attention-based neural network module, the Set Transformer, specifically designed to model interactions among elements in the input set. The model consists of an encoder and a decoder, both of which rely on attention mechanisms. In an effort to reduce computational complexity, we introduce an attention scheme inspired by inducing point methods from sparse Gaussian process literature. It reduces computation time of self-attention from quadratic to linear in the number of elements in the set. We show that our model is theoretically attractive and we evaluate it on a range of tasks, demonstrating increased performance compared to recent methods for set-structured data.

\*\*\*\*\*

#### MARGINALIZED AVERAGE ATTENTIONAL NETWORK FOR WEAKLY-SUPERVISED LEARNING

Yuan Yuan, Yueming Lyu, Xi Shen, Ivor W. Tsang, Dit-Yan Yeung

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 p

propose a fast algorithm to reduce the complexity of constructing MAA from  $\mathcal{O}(2^T)$  to  $\mathcal{O}(T^2)$ . 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

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.

\*\*\*\*\*

#### Probabilistic Program Induction for Intuitive Physics Game Play

Fahad Alhasoun

Recent findings suggest that humans deploy cognitive mechanism of physics simulation engines to simulate the physics of objects. We propose a framework for bots to deploy similar tools for interacting with intuitive physics environments. The framework employs a physics simulation in a probabilistic way to infer about moves performed by an agent in a setting governed by Newtonian laws of motion. However, methods of probabilistic programs can be slow in such setting due to their need to generate many samples. We complement the model with a model-free approach to aid the sampling procedures in becoming more efficient through learning from experience during game playing. We present an approach where a myriad of model-free approaches (a convolutional neural network in our model) and model-based approaches (probabilistic physics simulation) is able to achieve what neither could alone. This way the model outperforms an all model-free or all model-based approach. We discuss a case study showing empirical results of the performance of the model on the game of Flappy Bird.

\*\*\*\*\*

#### Stable Opponent Shaping in Differentiable Games

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, Shimon Whiteson

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

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.

\*\*\*\*\*

Deep Recurrent Gaussian Process with Variational Sparse Spectrum Approximation

Roman Föll, Bernard Haasdonk, Markus Hanselmann, Holger Ulmer

Modeling sequential data has become more and more important in practice. Some applications are autonomous driving, virtual sensors and weather forecasting. To model such systems, so called recurrent models are frequently used. In this paper we introduce several new Deep Recurrent Gaussian Process (DRGP) models based on the Sparse Spectrum Gaussian Process (SSGP) and the improved version, called Variational Sparse Spectrum Gaussian Process (VSSGP). We follow the recurrent structure given by an existing DRGP based on a specific variational sparse Nyström approximation, the recurrent Gaussian Process (RGP). Similar to previous work, we also variationally integrate out the input-space and hence can propagate uncertainty through the Gaussian Process (GP) layers. Our approach can deal with a larger class of covariance functions than the RGP, because its spectral nature allows variational integration in all stationary cases. Furthermore, we combine the (Variational) Sparse Spectrum ((V)SS) approximations with a well known inducing-input regularization framework. For the DRGP extension of these combined approximations and the simple (V)SS approximations an optimal variational distribution exists. We improve over current state of the art methods in prediction accuracy for experimental data-sets used for their evaluation and introduce a new data-set for engine control, named Emission.

\*\*\*\*\*

Understanding Composition of Word Embeddings via Tensor Decomposition

Abraham Frandsen, Rong Ge

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

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

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.

\*\*\*\*\*

SENSE: SEMANTICALLY ENHANCED NODE SEQUENCE EMBEDDING

Swati Rallapalli,Liang Ma,Mudhakar Srivatsa,Ananthram Swami,Heesung Kwon,Graham Bent,Christopher Simpkin

Effectively capturing graph node sequences in the form of vector embeddings is critical to many applications. We achieve this by (i) first learning vector embeddings of single graph nodes and (ii) then composing them to compactly represent node sequences. Specifically, we propose SENSE-S (Semantically Enhanced Node Sequence Embedding - for Single nodes), a skip-gram based novel embedding mechanism, for single graph nodes that co-learns graph structure as well as their textual descriptions. We demonstrate that SENSE-S vectors increase the accuracy of multi-label classification tasks by up to 50% and link-prediction tasks by up to 78% under a variety of scenarios using real datasets. Based on SENSE-S, we next propose generic SENSE to compute composite vectors that represent a sequence of nodes, where preserving the node order is important. We prove that this approach is efficient in embedding node sequences, and our experiments on real data confirm its high accuracy in node order decoding.

\*\*\*\*\*

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

Reinforcement learning agents need exploratory behaviors to escape from local optima. These behaviors may include both immediate dithering perturbation and temp

orally 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

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

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

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 w

ith 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.

\*\*\*\*\*

Learning to Drive by Observing the Best and Synthesizing the Worst  
Mayank Bansal, Alex Krizhevsky, Abhijit Ogale

Our goal is to train a policy for autonomous driving via imitation learning that is robust enough to drive a real vehicle. We find that standard behavior cloning is insufficient for handling complex driving scenarios, even when we leverage a perception system for preprocessing the input and a controller for executing the output on the car: 30 million examples are still not enough. We propose exposing the learner to synthesized data in the form of perturbations to the expert's driving, which creates interesting situations such as collisions and/or going off the road. Rather than purely imitating all data, we augment the imitation loss with additional losses that penalize undesirable events and encourage progress -- the perturbations then provide an important signal for these losses and lead to robustness of the learned model. We show that the model can handle complex situations in simulation, and present ablation experiments that emphasize the importance of each of our proposed changes and show that the model is responding to the appropriate causal factors. Finally, we demonstrate the model driving a car in the real world ( <https://sites.google.com/view/learn-to-drive> ).

\*\*\*\*\*

Adversarial Exploration Strategy for Self-Supervised Imitation Learning  
Zhang-Wei Hong, Tsu-Jui Fu, Tzu-Yun Shann, Yi-Hsiang Chang, Chun-Yi Lee

We present an adversarial exploration strategy, a simple yet effective imitation learning scheme that incentivizes exploration of an environment without any extrinsic reward or human demonstration. Our framework consists of a deep reinforcement learning (DRL) agent and an inverse dynamics model contesting with each other. The former collects training samples for the latter, and its objective is to maximize the error of the latter. The latter is trained with samples collected by the former, and generates rewards for the former when it fails to predict the actual action taken by the former. In such a competitive setting, the DRL agent learns to generate samples that the inverse dynamics model fails to predict correctly, and the inverse dynamics model learns to adapt to the challenging samples. We further propose a reward structure that ensures the DRL agent collects only moderately hard samples and not overly hard ones that prevent the inverse model from imitating effectively. We evaluate the effectiveness of our method on several OpenAI gym robotic arm and hand manipulation tasks against a number of baseline models. Experimental results show that our method is comparable to that directly trained with expert demonstrations, and superior to the other baselines even without any human priors.

\*\*\*\*\*

How Important is a Neuron

Kedar Dhamdhere, Mukund Sundararajan, Qiqi Yan

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 convolutional 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

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.

\*\*\*\*\*

Estimating Information Flow in DNNs

Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Brian Kingsbury, Igor Melnyk, Nam Nguyen, Yury Polyanskiy

We study the evolution of internal representations during deep neural network (DNN) training, aiming to demystify the compression aspect of the information bottleneck theory. The theory suggests that DNN training comprises a rapid fitting phase followed by a slower compression phase, in which the mutual information  $I(X;T)$  between the input  $X$  and internal representations  $T$  decreases. Several papers observe compression of estimated mutual information on different DNN models, but the true  $I(X;T)$  over these networks is provably either constant (discrete  $X$ ) or infinite (continuous  $X$ ). This work explains the discrepancy between theory and experiments, and clarifies what was actually measured by these past works. To this end, we introduce an auxiliary (noisy) DNN framework for which  $I(X;T)$  is a meaningful quantity that depends on the network's parameters. This noisy framework is shown to be a good proxy for the original (deterministic) DNN both in terms of performance and the learned representations. We then develop a rigorous estimator for  $I(X;T)$  in noisy DNNs and observe compression in various models. By relating  $I(X;T)$  in the noisy DNN to an information-theoretic communication problem, we show that compression is driven by the progressive clustering of hidden representations of inputs from the same class. Several methods to directly monitor clustering of hidden representations, both in noisy and deterministic DNNs, are used to show that meaningful clusters form in the  $T$  space. Finally, we return to the estimator of  $I(X;T)$  employed in past works, and demonstrate that while it fails to capture the true (vacuous) mutual information, it does serve as a measure for clustering. This clarifies the past observations of compression and isolates the geometric clustering of hidden representations as the true phenomenon of interest.

\*\*\*\*\*

Learning models for visual 3D localization with implicit mapping

Dan Rosenbaum, Frederic Besse, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami

We consider learning based methods for visual localization that do not require the construction of explicit maps in the form of point clouds or voxels. The goal is to learn an implicit representation of the environment at a higher, more abstract level, for instance that of objects. We propose to use a generative approach based on Generative Query Networks (GQNs, Eslami et al. 2018), asking the following questions: 1) Can GQN capture more complex scenes than those it was origi

nally demonstrated on? 2) Can GQN be used for localization in those scenes? To study this approach we consider procedurally generated Minecraft worlds, for which we can generate images of complex 3D scenes along with camera pose coordinates. We first show that GQNs, enhanced with a novel attention mechanism can capture the structure of 3D scenes in Minecraft, as evidenced by their samples. We then apply the models to the localization problem, comparing the results to a discriminative baseline, and comparing the ways each approach captures the task uncertainty.

\*\*\*\*\*

Hyper-Regularization: An Adaptive Choice for the Learning Rate in Gradient Descent

Guangzeng Xie, Hao Jin, Dachao Lin, Zhihua Zhang

We present a novel approach for adaptively selecting the learning rate in gradient descent methods. Specifically, we impose a regularization term on the learning rate via a generalized distance, and cast the joint updating process of the parameter and the learning rate into a maxmin problem. Some existing schemes such as AdaGrad (diagonal version) and WNGrad can be rederived from our approach. Based on our approach, the updating rules for the learning rate do not rely on the smoothness constant of optimization problems and are robust to the initial learning rate. We theoretically analyze our approach in full batch and online learning settings, which achieves comparable performances with other first-order gradient-based algorithms in terms of accuracy as well as convergence rate.

\*\*\*\*\*

Meta-Learning Update Rules for Unsupervised Representation Learning

Luke Metz, Niru Maheswaranathan, Brian Cheung, Jascha Sohl-Dickstein

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 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.

\*\*\*\*\*

Distributionally Robust Optimization Leads to Better Generalization: on SGD and Beyond

Jikai Hou, Kaixuan Huang, Zhihua Zhang

In this paper, we adopt distributionally robust optimization (DRO) (Ben-Tal et al., 2013) in hope to achieve a better generalization in deep learning tasks. We establish the generalization guarantees and analyze the localized Rademacher complexity for DRO, and conduct experiments to show that DRO obtains a better performance. We reveal the profound connection between SGD and DRO, i.e., selecting a batch can be viewed as choosing a distribution over the training set. From this perspective, we prove that SGD is prone to escape from bad stationary points and small batch SGD outperforms large batch SGD. We give an upper bound for the robust loss when SGD converges and keeps stable. We propose a novel Weighted SGD (WSGD) algorithm framework, which assigns high-variance weights to the data of the current batch. We devise a practical implement of WSGD that can directly optimize the robust loss. We test our algorithm on CIFAR-10 and CIFAR-100, and WSGD achieves significant improvements over the conventional SGD.

\*\*\*\*\*

## COLLABORATIVE MULTIAGENT REINFORCEMENT LEARNING IN HOMOGENEOUS SWARMS

Arbaaz Khan, Clark Zhang, Vijay Kumar, Alejandro Ribeiro

A deep reinforcement learning solution is developed for a collaborative multiagent system. Individual agents choose actions in response to the state of the environment, their own state, and possibly partial information about the state of other agents. Actions are chosen to maximize a collaborative long term discounted reward that encompasses the individual rewards collected by each agent. The paper focuses on developing a scalable approach that applies to large swarms of homogeneous agents. This is accomplished by forcing the policies of all agents to be the same resulting in a constrained formulation in which the experiences of each agent inform the learning process of the whole team, thereby enhancing the sample efficiency of the learning process. A projected coordinate policy gradient descent algorithm is derived to solve the constrained reinforcement learning problem. Experimental evaluations in collaborative navigation, a multi-predator-multi-prey game, and a multiagent survival game show marked improvements relative to methods that do not exploit the policy equivalence that naturally arises in homogeneous swarms.

\*\*\*\*\*

## Deconfounding Reinforcement Learning in Observational Settings

Chaochao Lu, José Miguel Hernández Lobato

In this paper, we propose a general formulation to cope with a family of reinforcement learning tasks in observational settings, that is, learning good policies solely from the historical data produced by real environments with confounders (i.e., the factors affecting both actions and rewards). Based on the proposed approach, we extend one representative of reinforcement learning algorithms: the Actor-Critic method, to its deconfounding variant, which is also straightforward to be applied to other algorithms. In addition, due to lack of datasets in this direction, a benchmark is developed for deconfounding reinforcement learning algorithms by revising OpenAI Gym and MNIST. We demonstrate that the proposed algorithms are superior to traditional reinforcement learning algorithms in confounded environments. To the best of our knowledge, this is the first time that confounders are taken into consideration for addressing full reinforcement learning problems.

\*\*\*\*\*

## Emergent Coordination Through Competition

Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess, Thore Graepel

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

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

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.

\*\*\*\*\*

#### Backdrop: Stochastic Backpropagation

Siavash Golkar, Kyle Cranmer

We introduce backdrop, a flexible and simple-to-implement method, intuitively described as dropout acting only along the backpropagation pipeline. Backdrop is implemented via one or more masking layers which are inserted at specific points along the network. Each backdrop masking layer acts as the identity in the forward pass, but randomly masks parts of the backward gradient propagation. Intuitively, inserting a backdrop layer after any convolutional layer leads to stochastic gradients corresponding to features of that scale. Therefore, backdrop is well suited for problems in which the data have a multi-scale, hierarchical structure. Backdrop can also be applied to problems with non-decomposable loss functions where standard SGD methods are not well suited. We perform a number of experiments and demonstrate that backdrop leads to significant improvements in generalization.

\*\*\*\*\*

#### Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization

Hesham Mostafa, Xin Wang

Modern deep neural networks are highly overparameterized, and often of huge sizes. A number of post-training model compression techniques, such as distillation, pruning and quantization, can reduce the size of network parameters by a substantial fraction with little loss in performance. However, training a small network of the post-compression size de novo typically fails to reach the same level of accuracy achieved by compression of a large network, leading to a widely-held belief that gross overparameterization is essential to effective learning. In this work, we argue that this is not necessarily true. We describe a dynamic sparse reparameterization technique that closed the performance gap between a model compressed through iterative pruning and a model of the post-compression size trained de novo. We applied our method to training deep residual networks and showed that it outperformed existing reparameterization techniques, yielding the best accuracy for a given parameter budget for training. Compared to existing

dynamic reparameterization methods that reallocate non-zero parameters during training, our approach achieved better performance at lower computational cost. Our method is not only of practical value for training under stringent memory constraints, but also potentially informative to theoretical understanding of generalization properties of overparameterized deep neural networks.

\*\*\*\*\*

#### Structured Neural Summarization

Patrick Fernandes, Miltiadis Allamanis, Marc Brockschmidt

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

Çaglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, Nando de Freitas

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.

\*\*\*\*\*

#### An Efficient and Margin-Approaching Zero-Confidence Adversarial Attack

Yang Zhang, Shiyu Chang, Mo Yu, Kaizhi Qian

There are two major paradigms of white-box adversarial attacks that attempt to impose input perturbations. The first paradigm, called the fix-perturbation attack, crafts adversarial samples within a given perturbation level. The second paradigm, called the zero-confidence attack, finds the smallest perturbation needed to cause misclassification, also known as the margin of an input feature. While the former paradigm is well-resolved, the latter is not. Existing zero-confidence attacks either introduce significant approximation errors, or are too time-consuming. We therefore propose MarginAttack, a zero-confidence attack framework that is able to compute the margin with improved accuracy and efficiency. Our experiments show that MarginAttack is able to compute a smaller margin than the state-of-the-art zero-confidence attacks, and matches the state-of-the-art fix-perturbation attacks. In addition, it runs significantly faster than the Carlini-Wagner attack, currently the most accurate zero-confidence attack algorithm.

\*\*\*\*\*

#### In Your Pace: Learning the Right Example at the Right Time

Guy Hacohen, Daphna Weinshall

Training neural networks is traditionally done by sequentially providing random mini-batches sampled uniformly from the entire dataset. In our work, we show that sampling mini-batches non-uniformly can both enhance the speed of learning and improve the final accuracy of the trained network. Specifically, we decompose the problem using the principles of curriculum learning: first, we sort the data by some difficulty measure; second, we sample mini-batches with a gradually increasing

easing level of difficulty. We focus on CNNs trained on image recognition. Initially, we define the difficulty of a training image using transfer learning from some competitive "teacher" network trained on the Imagenet database, showing improvement in learning speed and final performance for both small and competitive networks, using the CIFAR-10 and the CIFAR-100 datasets. We then suggest a bootstrap alternative to evaluate the difficulty of points using the same network without relying on a "teacher" network, thus increasing the applicability of our suggested method. We compare this approach to a related version of Self-Paced Learning, showing that our method benefits learning while SPL impairs it.

\*\*\*\*\*

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

Min Liu, Fupin Yao, Chiho Choi, Ayan Sinha, Karthik Ramani

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

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.

\*\*\*\*\*

A Case for Object Compositionality in Deep Generative Models of Images

Sjoerd van Steenkiste, Karol Kurach, Sylvain Gelly

Deep generative models seek to recover the process with which the observed data was generated. They may be used to synthesize new samples or to subsequently extract representations. Successful approaches in the domain of images are driven by several core inductive biases. However, a bias to account for the compositional way in which humans structure a visual scene in terms of objects has frequently been overlooked. In this work we propose to structure the generator of a GAN to consider objects and their relations explicitly, and generate images by means of composition. This provides a way to efficiently learn a more accurate generative model of real-world images, and serves as an initial step towards learning corresponding object representations. We evaluate our approach on several multi-object image datasets, and find that the generator learns to identify and disentangle information corresponding to different objects at a representational level.

A human study reveals that the resulting generative model is better at generating images that are more faithful to the reference distribution.

\*\*\*\*\*

SEGEN: SAMPLE-ENSEMBLE GENETIC EVOLUTIONARY NETWORK MODEL

Jiawei Zhang, Limeng Cui, Fisher B. Gouza

Deep learning, a rebranding of deep neural network research works, has achieved

a remarkable success in recent years. With multiple hidden layers, deep learning models aim at computing the hierarchical feature representations of the observational data. Meanwhile, due to its severe disadvantages in data consumption, computational resources, parameter tuning costs and the lack of result explainability, deep learning has also suffered from lots of criticism. In this paper, we will introduce a new representation learning model, namely "Sample-Ensemble Genetic Evolutionary Network" (SEGEN), which can serve as an alternative approach to deep learning models. Instead of building one single deep model, based on a set of sampled sub-instances, SEGEN adopts a genetic-evolutionary learning strategy to build a group of unit models generations by generations. The unit models incorporated in SEGEN can be either traditional machine learning models or the recent deep learning models with a much "narrower" and "shallower" architecture. The learning results of each instance at the final generation will be effectively combined from each unit model via diffusive propagation and ensemble learning strategies. From the computational perspective, SEGEN requires far less data, fewer computational resources and parameter tuning efforts, but has sound theoretic interpretability of the learning process and results. Extensive experiments have been done on several different real-world benchmark datasets, and the experimental results obtained by SEGEN have demonstrated its advantages over the state-of-the-art representation learning models.

\*\*\*\*\*

#### Relational Forward Models for Multi-Agent Learning

Andrea Tacchetti, H. Francis Song, Pedro A. M. Mediano, Vinicius Zambaldi, János Krašár, Neil C. Rabinowitz, Thore Graepel, Matthew Botvinick, Peter W. Battaglia

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

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

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 network. 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

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 (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 (Our code is available at <https://github.com/bhargavl04/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.

\*\*\*\*\*

The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Minima and Regularization Effects

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, Jinwen Ma

Understanding the behavior of stochastic gradient descent (SGD) in the context of deep neural networks has raised lots of concerns recently. Along this line, we theoretically study a general form of gradient based optimization dynamics with unbiased noise, which unifies SGD and standard Langevin dynamics. Through investigating this general optimization dynamics, we analyze the behavior of SGD on escaping from minima and its regularization effects. A novel indicator is derived to characterize the efficiency of escaping from minima through measuring the alignment of noise covariance and the curvature of loss function. Based on this indicator, two conditions are established to show which type of noise structure is superior to isotropic noise in term of escaping efficiency. We further show that the anisotropic noise in SGD satisfies the two conditions, and thus helps to escape from sharp and poor minima effectively, towards more stable and flat minima that typically generalize well. We verify our understanding through comparing

this anisotropic diffusion with full gradient descent plus isotropic diffusion (i.e. Langevin dynamics) and other types of position-dependent noise.

\*\*\*\*\*

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



ys

Quynh Nguyen, Mahesh Chandra Mukkamala, Matthias Hein

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.

\*\*\*\*\*

Representation-Constrained Autoencoders and an Application to Wireless Positioning

Pengzhi Huang, Emre Gonultas, Said Medjkouh, Oscar Castaneda, Olav Tirkkonen, Tom Goldstein, Christoph Studer

In a number of practical applications that rely on dimensionality reduction, the dataset or measurement process provides valuable side information that can be incorporated when learning low-dimensional embeddings. We propose the inclusion of pairwise representation constraints into autoencoders (AEs) with the goal of promoting application-specific structure. We use synthetic results to show that only a small amount of AE representation constraints are required to substantially improve the local and global neighborhood preserving properties of the learned embeddings. To demonstrate the efficacy of our approach and to illustrate a practical application that naturally provides such representation constraints, we focus on wireless positioning using a recently proposed channel charting framework. We show that representation-constrained AEs recover the global geometry of the learned low-dimensional representations, which enables channel charting to perform approximate positioning without access to global navigation satellite systems or supervised learning methods that rely on extensive measurement campaigns.

\*\*\*\*\*

Backprop with Approximate Activations for Memory-efficient Network Training

Ayan Chakrabarti, Benjamin Moseley

With innovations in architecture design, deeper and wider neural network models deliver improved performance on a diverse variety of tasks. But the increased memory footprint of these models presents a challenge during training, when all intermediate layer activations need to be stored for back-propagation. Limited GPU memory forces practitioners to make sub-optimal choices: either train inefficiently with smaller batches of examples; or limit the architecture to have lower depth and width, and fewer layers at higher spatial resolutions. This work introduces an approximation strategy that significantly reduces a network's memory footprint during training, but has negligible effect on training performance and computational expense. During the forward pass, we replace activations with lower-precision approximations immediately after they have been used by subsequent layers, thus freeing up memory. The approximate activations are then used during the backward pass. This approach limits the accumulation of errors across the forward and backward pass---because the forward computation across the network still happens at full precision, and the approximation has a limited effect when computing gradients to a layer's input. Experiments, on CIFAR and ImageNet, show that using our approach with 8- and even 4-bit fixed-point approximations of 32-bit floating-point activations has only a minor effect on training and validation performance, while affording significant savings in memory usage.

\*\*\*\*\*

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

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 at 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.

\*\*\*\*\*

#### Auto-Encoding Knockoff Generator for FDR Controlled Variable Selection

Ying Liu, Cheng Zheng

A new statistical procedure (Candes, 2018) has provided a way to identify important factors using any supervised learning method controlling for FDR. This line of research has shown great potential to expand the horizon of machine learning methods beyond the task of prediction, to serve the broader need for scientific researches for interpretable findings. However, the lack of a practical and flexible method to generate knockoffs remains the major obstacle for wide application of Model-X procedure. This paper fills in the gap by proposing a model-free knockoff generator which approximates the correlation structure between features through latent variable representation. We demonstrate our proposed method can achieve FDR control and better power than two existing methods in various simulated settings and a real data example for finding mutations associated with drug resistance in HIV-1 patients.

\*\*\*\*\*

#### Deep Convolutional Networks as shallow Gaussian Processes

Adrià Garriga-Alonso, Carl Edward Rasmussen, Laurence Aitchison

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

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 forms 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.

\*\*\*\*\*

#### A Biologically Inspired Visual Working Memory for Deep Networks

Ethan Harris, Mahesan Niranjan, Jonathon Hare

The ability to look multiple times through a series of pose-adjusted glimpses is fundamental to human vision. This critical faculty allows us to understand highly complex visual scenes. Short term memory plays an integral role in aggregating the information obtained from these glimpses and informing our interpretation of the scene. Computational models have attempted to address glimpsing and visual attention but have failed to incorporate the notion of memory. We introduce a novel, biologically inspired visual working memory architecture that we term the Hebb-Rosenblatt memory. We subsequently introduce a fully differentiable Short Term Attentive Working Memory model (STAWM) which uses transformational attention to learn a memory over each image it sees. The state of our Hebb-Rosenblatt memory is embedded in STAWM as the weights space of a layer. By projecting different queries through this layer we can obtain goal-oriented latent representations for tasks including classification and visual reconstruction. Our model obtains highly competitive classification performance on MNIST and CIFAR-10. As demonstrated through the CelebA dataset, to perform reconstruction the model learns to make a sequence of updates to a canvas which constitute a parts-based representation. Classification with the self supervised representation obtained from MNIST is shown to be in line with the state of the art models (none of which use a visual attention mechanism). Finally, we show that STAWM can be trained under the dual constraints of classification and reconstruction to provide an interpretable visual sketchpad which helps open the 'black-box' of deep learning.

\*\*\*\*\*

#### On Generalization Bounds of a Family of Recurrent Neural Networks

Minshuo Chen, Xingguo Li, Tuo Zhao

Recurrent Neural Networks (RNNs) have been widely applied to sequential data analysis. Due to their complicated modeling structures, however, the theory behind is still largely missing. To connect theory and practice, we study the generalization properties of vanilla RNNs as well as their variants, including Minimal Gated Unit (MGU) and Long Short Term Memory (LSTM) RNNs. Specifically, our theory is established under the PAC-Learning framework. The generalization bound is presented in terms of the spectral norms of the weight matrices and the total number of parameters. We also establish refined generalization bounds with additional norm assumptions, and draw a comparison among these bounds. We remark: (1) Our generalization bound for vanilla RNNs is significantly tighter than the best of existing results; (2) We are not aware of any other generalization bounds for MGU and LSTM in the existing literature; (3) We demonstrate the advantages of these variants in generalization.

\*\*\*\*\*

#### The Cakewalk Method

Uri Patish, Shimon Ullman

Combinatorial optimization is a common theme in computer science. While in general such problems are NP-Hard, from a practical point of view, locally optimal solutions can be useful. In some combinatorial problems however, it can be hard to define meaningful solution neighborhoods that connect large portions of the search space, thus hindering methods that search this space directly. We suggest to circumvent such cases by utilizing a policy gradient algorithm that transforms the problem to the continuous domain, and to optimize a new surrogate objective that renders the former as generic stochastic optimizer. This is achieved by producing a surrogate objective whose distribution is fixed and predetermined, thus removing the need to fine-tune various hyper-parameters in a case by case manner. Since we are interested in methods which can successfully recover locally optimal solutions, we use the problem of finding locally maximal cliques as a challenging experimental benchmark, and we report results on a large dataset of graph

s that is designed to test clique finding algorithms. Notably, we show in this benchmark that fixing the distribution of the surrogate is key to consistently recovering locally optimal solutions, and that our surrogate objective leads to an algorithm that outperforms other methods we have tested in a number of measures

\*\*\*\*\*

#### Adaptive Estimators Show Information Compression in Deep Neural Networks

Ivan Chelombiev, Conor Houghton, Cian O'Donnell

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. Secondly, 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.

\*\*\*\*\*

#### Neural Causal Discovery with Learnable Input Noise

Tailin Wu, Thomas Breuel, Jan Kautz

Learning causal relations from observational time series with nonlinear interactions and complex causal structures is a key component of human intelligence, and has a wide range of applications. Although neural nets have demonstrated their effectiveness in a variety of fields, their application in learning causal relations has been scarce. This is due to both a lack of theoretical results connecting risk minimization and causality (enabling function approximators like neural nets to apply), and a lack of scalability in prior causal measures to allow for expressive function approximators like neural nets to apply. In this work, we propose a novel causal measure and algorithm using risk minimization to infer causal relations from time series. We demonstrate the effectiveness and scalability of our algorithms to learn nonlinear causal models in synthetic datasets as compared to other methods, and its effectiveness in inferring causal relations in a video game environment and real-world heart-rate vs. breath-rate and rat brain EEG datasets.

\*\*\*\*\*

#### Dataset Distillation

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, Alexei A. Efros

Model distillation aims to distill the knowledge of a complex model into a simpler one. In this paper, we consider an alternative formulation called *dataset distillation*: we keep the model fixed and instead attempt to distill the knowledge from a large training dataset into a small one. The idea is to synthesize a small number of data points that do not need to come from the correct data distribution, but will, when given to the learning algorithm as training data, approximate the model trained on the original data. For example, we show that it is possible to compress \$60,000\$ MNIST training images into just \$10\$ synthetic *distilled images* (one per class) and achieve close to original performance with only a few steps of gradient descent, given a particular fixed network initialization. We evaluate our method in a wide range of initialization settings and with different learning objectives. Experiments on multiple datasets show the advantage of our approach compared to alternative methods in most settings.

gs.

\*\*\*\*\*

#### Gradient-based learning for F-measure and other performance metrics

Yu Gai,Zheng Zhang,Kyunghyun Cho

Many important classification performance metrics, e.g.  $F_1$ -measure, are non-differentiable and non-decomposable, and are thus unfriendly to gradient descent algorithm.

Consequently, despite their popularity as evaluation metrics, these metrics are rarely optimized as training objectives in neural network community.

In this paper, we propose an empirical utility maximization scheme with provable learning guarantees to address the non-differentiability of these metrics.

We then derive a strongly consistent gradient estimator to handle non-decomposability.

These innovations enable end-to-end optimization of these metrics with the same computational complexity as optimizing a decomposable and differentiable metric, e.g. cross-entropy loss.

\*\*\*\*\*

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

Florian Mai,Lukas Galke,Ansgar Scherp

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 Continuous 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%.

\*\*\*\*\*

#### Using Deep Siamese Neural Networks to Speed up Natural Products Research

Nicholas Roberts,Poornav S. Purushothama,Vishal T. Vasudevan,Siddarth Ravichandran,Chen Zhang,William H. Gerwick,Garrison W. Cottrell

Natural products (NPs, compounds derived from plants and animals) are an important source of novel disease treatments. A bottleneck in the search for new NPs is structure determination. One method is to use 2D Nuclear Magnetic Resonance (NMR) imaging, which indicates bonds between nuclei in the compound, and hence is the "fingerprint" of the compound. Computing a similarity score between 2D NMR spectra for a novel compound and a compound whose structure is known helps determine the structure of the novel compound. Standard approaches to this problem do not appear to scale to larger databases of compounds. Here we use deep convolutional Siamese networks to map NMR spectra to a cluster space, where similarity is given by the distance in the space. This approach results in an AUC score that is more than four times better than an approach using Latent Dirichlet Allocation.

\*\*\*\*\*

#### Radial Basis Feature Transformation to Arm CNNs Against Adversarial Attacks

Saeid Asgari Taghanaki,Shekoofeh Azizi,Ghassan Hamarneh

The linear and non-flexible nature of deep convolutional models makes them vulnerable to carefully crafted adversarial perturbations. To tackle this problem, in this paper, we propose a nonlinear radial basis convolutional feature transformation by learning the Mahalanobis distance function that maps the input convolutional features from the same class into tight clusters. In such a space, the clu

sters become compact and well-separated, which prevent small adversarial perturbations from forcing a sample to cross the decision boundary. We test the proposed method on three publicly available image classification and segmentation datasets namely, MNIST, ISBI ISIC skin lesion, and NIH ChestX-ray14. We evaluate the robustness of our method to different gradient (targeted and untargeted) and non-gradient based attacks and compare it to several non-gradient masking defense strategies. Our results demonstrate that the proposed method can boost the performance of deep convolutional neural networks against adversarial perturbations without accuracy drop on clean data.

\*\*\*\*\*

Fluctuation-dissipation relations for stochastic gradient descent

Sho Yaida

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

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.

\*\*\*\*\*

Probabilistic Binary Neural Networks

Jorn W.T. Peters, Tim Genewein, Max Welling

Low bit-width weights and activations are an effective way of combating the increasing need for both memory and compute power of Deep Neural Networks. In this work, we present a probabilistic training method for Neural Network with both binary weights and activations, called PBNNet. By embracing stochasticity during training, we circumvent the need to approximate the gradient of functions for which the derivative is zero almost always, such as  $\text{sign}(\cdot)$ , while still obtaining a fully Binary Neural Network at test time. Moreover, it allows for anytime ensemble predictions for improved performance and uncertainty estimates by sampling from the weight distribution. Since all operations in a layer of the PBNNet operate on random variables, we introduce stochastic versions of Batch Normalization and max pooling, which transfer well to a deterministic network at test time. We evaluate two related training methods for the PBNNet: one in which activation distributions are propagated throughout the network, and one in which binary activations are sampled in each layer. Our experiments indicate that sampling the binary activations is an important element for stochastic training of binary Neural Networks.

\*\*\*\*\*

TabNN: A Universal Neural Network Solution for Tabular Data

Guolin Ke, Jia Zhang, Zhenhui Xu, Jiang Bian, Tie-Yan Liu

Neural Network (NN) has achieved state-of-the-art performances in many tasks within image, speech, and text domains. Such great success is mainly due to special

structure design to fit the particular data patterns, such as CNN capturing spatial locality and RNN modeling sequential dependency. Essentially, these specific NNs achieve good performance by leveraging the prior knowledge over corresponding domain data. Nevertheless, there are many applications with all kinds of tabular data in other domains. Since there are no shared patterns among these diverse tabular data, it is hard to design specific structures to fit them all. Without careful architecture design based on domain knowledge, it is quite challenging for NN to reach satisfactory performance in these tabular data domains. To fill the gap of NN in tabular data learning, we propose a universal neural network solution, called TabNN, to derive effective NN architectures for tabular data in all kinds of tasks automatically. Specifically, the design of TabNN follows two principles: \emph{to explicitly leverages expressive feature combinations} and \emph{to reduce model complexity}. Since GBDT has empirically proven its strength in modeling tabular data, we use GBDT to power the implementation of TabNN. Comprehensive experimental analysis on a variety of tabular datasets demonstrate that TabNN can achieve much better performance than many baseline solutions.

\*\*\*\*\*

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

Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbach, Karsten Borgwardt

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.

\*\*\*\*\*

Hierarchically Clustered Representation Learning

Su-Jin Shin, Kyungwoo Song, Il-Chul Moon

The joint optimization of representation learning and clustering in the embedding space has experienced a breakthrough in recent years. In spite of the advance, clustering with representation learning has been limited to flat-level categories, which oftentimes involves cohesive clustering with a focus on instance relations. To overcome the limitations of flat clustering, we introduce hierarchically clustered representation learning (HCRL), which simultaneously optimizes representation learning and hierarchical clustering in the embedding space. Specifically, we place a nonparametric Bayesian prior on embeddings to handle dynamic mixture hierarchies under the variational autoencoder framework, and to adopt the generative process of a hierarchical-versioned Gaussian mixture model. Compared with a few prior works focusing on unifying representation learning and hierarchical clustering, HCRL is the first model to consider a generation of deep embeddings from every component of the hierarchy, not just leaf components. This generation process enables more meaningful separations and mergers of clusters via branches in a hierarchy. In addition to obtaining hierarchically clustered embeddings, we can reconstruct data by the various abstraction levels, infer the intrinsic hierarchical structure, and learn the level-proportion features. We conducted evaluations with image and text domains, and our quantitative analyses showed competent likelihoods and the best accuracies compared with the baselines.

\*\*\*\*\*

Detecting Egregious Responses in Neural Sequence-to-sequence Models

Tianxing He, James Glass

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 efficient

ly. 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

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 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

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.

\*\*\*\*\*

## Point Cloud GAN

Chun-Liang Li, Manzil Zaheer, Yang Zhang, Barnabás Póczos, Ruslan Salakhutdinov

Generative Adversarial Networks (GAN) can achieve promising performance on learning complex data distributions on different types of data. In this paper, we first show that a straightforward extension of an existing GAN algorithm is not applicable to point clouds, because the constraint required for discriminators is undefined for set data. We propose a two fold modification to a GAN algorithm to be able to generate point clouds (PC-GAN). First, we combine ideas from hierarchical Bayesian modeling and implicit generative models by learning a hierarchical and interpretable sampling process. A key component of our method is that we train a posterior inference network for the hidden variables. Second, PC-GAN defines a generic framework that can incorporate many existing GAN algorithms. We further propose a sandwiching objective, which results in a tighter Wasserstein distance estimate than the commonly used dual form in WGAN. We validate our claims on the ModelNet40 benchmark dataset and observe that PC-GAN trained by the sandwiching objective achieves better results on test data than existing methods. We also conduct studies on several tasks, including generalization on unseen point clouds, latent space interpolation, classification, and image to point clouds



ransformation, to demonstrate the versatility of the proposed PC-GAN algorithm.

\*\*\*\*\*

#### Analysing Mathematical Reasoning Abilities of Neural Models

David Saxton, Edward Grefenstette, Felix Hill, Pushmeet Kohli

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 splits 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.

\*\*\*\*\*

#### Learning Information Propagation in the Dynamical Systems via Information Bottleneck Hierarchy

Gaurav Gupta, Mohamed Ridha Znaidi, Paul Bogdan

Extracting relevant information, causally inferring and predicting the future states with high accuracy is a crucial task for modeling complex systems. The endeavor to address these tasks is made even more challenging when we have to deal with high-dimensional heterogeneous data streams. Such data streams often have higher-order inter-dependencies across spatial and temporal dimensions. We propose to perform a soft-clustering of the data and learn its dynamics to produce a compact dynamical model while still ensuring the original objectives of causal inference and accurate predictions. To efficiently and rigorously process the dynamics of soft-clustering, we advocate for an information theory inspired approach that incorporates stochastic calculus and seeks to determine a trade-off between the predictive accuracy and compactness of the mathematical representation. We cast the model construction as a maximization of the compression of the state variables such that the predictive ability and causal interdependence (relatedness) constraints between the original data streams and the compact model are closely bounded. We provide theoretical guarantees concerning the convergence of the proposed learning algorithm. To further test the proposed framework, we consider a high-dimensional Gaussian case study and describe an iterative scheme for updating the new model parameters. Using numerical experiments, we demonstrate the benefits on compression and prediction accuracy for a class of dynamical systems. Finally, we apply the proposed algorithm to the real-world dataset of multimodal sentiment intensity and show improvements in prediction with reduced dimensions.

\*\*\*\*\*

#### Improving Generative Adversarial Imitation Learning with Non-expert Demonstrations

Voot Tangkaratt, Masashi Sugiyama

Imitation learning aims to learn an optimal policy from expert demonstrations and its recent combination with deep learning has shown impressive performance. However, collecting a large number of expert demonstrations for deep learning is time-consuming and requires much expert effort. In this paper, we propose a method to improve generative adversarial imitation learning by using additional information from non-expert demonstrations which are easier to obtain. The key idea of our method is to perform multiclass classification to learn discriminator functions where non-expert demonstrations are regarded as being drawn from an extra class. Experiments in continuous control tasks demonstrate that our method learn

s better policies than the generative adversarial imitation learning baseline when the number of expert demonstrations is small.

\*\*\*\*\*

Ordered Neurons: Integrating Tree Structures into Recurrent Neural Networks

Yikang Shen, Shawn Tan, Alessandro Sordoni, Aaron Courville

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

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.

\*\*\*\*\*

MixFeat: Mix Feature in Latent Space Learns Discriminative Space

Yoichi Yaguchi, Fumiyuki Shiratani, Hidekazu Iwaki

Deep learning methods perform well in various tasks. However, the over-fitting problem, which causes the performance to decrease for unknown data, remains. We hence propose a method named MixFeat that directly creates latent spaces in a network that can distinguish classes. MixFeat mixes two feature maps in each latent space in the network and uses unmixed labels for learning. We discuss the difference between a method that mixes only features (MixFeat) and a method that mixes both features and labels (mixup and its family). Mixing features repeatedly is effective in expanding feature diversity, but mixing labels repeatedly makes learning difficult. MixFeat makes it possible to obtain the advantages of repeated mixing by mixing only features. We report improved results obtained using existing network models with MixFeat on CIFAR-10/100 datasets. In addition, we show that MixFeat effectively reduces the over-fitting problem even when the training dataset is small or contains errors. MixFeat is easy to implement and can be added to various network models without additional computational cost in the inference phase.

\*\*\*\*\*

Understanding the Asymptotic Performance of Model-Based RL Methods

William Whitney, Rob Fergus

In complex simulated environments, model-based reinforcement learning methods typically lag the asymptotic performance of model-free approaches. This paper uses two MuJoCo environments to understand this gap through a series of ablation experiments designed to separate the contributions of the dynamics model and planner. These reveal the importance of long planning horizons, beyond those typically used. A dynamics model that directly predicts distant states, based on current state and a long sequence of actions, is introduced. This avoids the need for ma

ny recursions during long-range planning, and thus is able to yield more accurate state estimates. These accurate predictions allow us to uncover the relationship between model accuracy and performance, and translate to higher task reward that matches or exceeds current state-of-the-art model-free approaches.

\*\*\*\*\*

#### Weakly-supervised Knowledge Graph Alignment with Adversarial Learning

Meng Qu, Jian Tang, Yoshua Bengio

Aligning knowledge graphs from different sources or languages, which aims to align both the entity and relation, is critical to a variety of applications such as knowledge graph construction and question answering. Existing methods of knowledge graph alignment usually rely on a large number of aligned knowledge triplets to train effective models. However, these aligned triplets may not be available or are expensive to obtain for many domains. Therefore, in this paper we study how to design fully-unsupervised methods or weakly-supervised methods, i.e., to align knowledge graphs without or with only a few aligned triplets. We propose an unsupervised framework based on adversarial training, which is able to map the entities and relations in a source knowledge graph to those in a target knowledge graph. This framework can be further seamlessly integrated with existing supervised methods, where only a limited number of aligned triplets are utilized as guidance. Experiments on real-world datasets prove the effectiveness of our proposed approach in both the weakly-supervised and unsupervised settings.

\*\*\*\*\*

#### Learning Neural Random Fields with Inclusive Auxiliary Generators

Yunfu Song, Zhijian Ou

Neural random fields (NRFs), which are defined by using neural networks to implement potential functions in undirected models, provide an interesting family of model spaces for machine learning. In this paper we develop a new approach to learning NRFs with inclusive-divergence minimized auxiliary generator - the inclusive-NRF approach, for continuous data (e.g. images), with solid theoretical examination on exploiting gradient information in model sampling. We show that inclusive-NRFs can be flexibly used in unsupervised/supervised image generation and semi-supervised classification, and empirically to the best of our knowledge, represent the best-performed random fields in these tasks. Particularly, inclusive-NRFs achieve state-of-the-art sample generation quality on CIFAR-10 in both unsupervised and supervised settings. Semi-supervised inclusive-NRFs show strong classification results on par with state-of-the-art generative model based semi-supervised learning methods, and simultaneously achieve superior generation, on the widely benchmarked datasets - MNIST, SVHN and CIFAR-10.

\*\*\*\*\*

#### Countering Language Drift via Grounding

Jason Lee, Kyunghyun Cho, Douwe Kiela

While reinforcement learning (RL) shows a lot of promise for natural language processing—e.g. when fine-tuning natural language systems for optimizing a certain objective—there has been little investigation into potential language drift: when an external reward is used to train a system, the agents' communication protocol may easily and radically diverge from natural language. By re-casting translation as a communication game, we show that language drift indeed happens when pre-trained agents are fine-tuned with policy gradient methods. We contend that simply adding a "naturalness" constraint to the reward, e.g. by using language model log likelihood, does not fully address the issue, and argue that (perceptual) grounding is required. That is, while language model constraints impose syntactic conformity, they do not lead to semantic correspondence. Our experiments show that grounded models give the best communication performance, while retaining English syntax along with the ability to convey the intended semantics.

\*\*\*\*\*

#### Robustness and Equivariance of Neural Networks

Amit Deshpande, Sandesh Kamath, K.V. Subrahmanyam

Neural networks models are known to be vulnerable to geometric transformations as well as small pixel-wise perturbations of input. Convolutional Neural Networks

(CNNs) are translation-equivariant but can be easily fooled using rotations and small pixel-wise perturbations. Moreover, CNNs require sufficient translations in

their training data to achieve translation-invariance. Recent work by Cohen & Welling (2016), Worrall et al. (2016), Kondor & Trivedi (2018), Cohen & Welling (2017), Marcos et al. (2017), and Esteves et al. (2018) has gone beyond translations,

and constructed rotation-equivariant or more general group-equivariant neural network models. In this paper, we do an extensive empirical study of various

rotation-equivariant neural network models to understand how effectively they learn rotations. This includes Group-equivariant Convolutional Networks (GCNNs) by Cohen & Welling (2016), Harmonic Networks (H-Nets) by Worrall et al.

(2016), Polar Transformer Networks (PTN) by Esteves et al. (2018) and Rotation equivariant vector field networks by Marcos et al. (2017). We empirically compare

the ability of these networks to learn rotations efficiently in terms of their number of parameters, sample complexity, rotation augmentation used in training. We compare them against each other as well as Standard CNNs. We observe that as these rotation-equivariant neural networks learn rotations, they instead become

more vulnerable to small pixel-wise adversarial attacks, e.g., Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), in comparison with Standard CNNs. In other words, robustness to geometric transformations in these models comes at the cost of robustness to small pixel-wise perturbations.

\*\*\*\*\*

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

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

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.

\*\*\*\*\*

#### Progressive Weight Pruning Of Deep Neural Networks Using ADMM

Shaokai Ye, Tianyun Zhang, Kaiqi Zhang, Jiayu Li, Kaidi Xu, Yunfei Yang, Fuxun Yu, Jiang Tang, Makan Fardad, Sijia Liu, Xiang Chen, Xue Lin, Yanzhi Wang

Deep neural networks (DNNs) although achieving human-level performance in many domains, have very large model size that hinders their broader applications on edge computing devices. Extensive research work have been conducted on DNN model compression or pruning. However, most of the previous work took heuristic approaches. This work proposes a progressive weight pruning approach based on ADMM (Alternating Direction Method of Multipliers), a powerful technique to deal with non-convex optimization problems with potentially combinatorial constraints. Motivated by dynamic programming, the proposed method reaches extremely high pruning rate by using partial prunings with moderate pruning rates. Therefore, it resolves the accuracy degradation and long convergence time problems when pursuing extremely high pruning ratios. It achieves up to 34× pruning rate for ImageNet dataset and 167× pruning rate for MNIST dataset, significantly higher than those reached by the literature work. Under the same number of epochs, the proposed method also achieves faster convergence and higher compression rates. The codes and pruned DNN models are released in the anonymous link [bit.ly/2zxdlls](https://bit.ly/2zxdlls).

\*\*\*\*\*

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

Andrew Brock, Jeff Donahue, Karen Simonyan

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

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 na

tural 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 dataset. 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

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.

\*\*\*\*\*

#### The Importance of Norm Regularization in Linear Graph Embedding: Theoretical Analysis and Empirical Demonstration

Yihan Gao, Chao Zhang, Jian Peng, Aditya Parameswaran

Learning distributed representations for nodes in graphs is a crucial primitive in network analysis with a wide spectrum of applications. Linear graph embedding methods learn such representations by optimizing the likelihood of both positive and negative edges while constraining the dimension of the embedding vectors. We argue that the generalization performance of these methods is not due to the dimensionality constraint as commonly believed, but rather the small norm of embedding vectors. Both theoretical and empirical evidence are provided to support this argument: (a) we prove that the generalization error of these methods can be bounded by limiting the norm of vectors, regardless of the embedding dimension; (b) we show that the generalization performance of linear graph embedding methods is correlated with the norm of embedding vectors, which is small due to the early stopping of SGD and the vanishing gradients. We performed extensive experiments to validate our analysis and showcased the importance of proper norm regularization in practice.

\*\*\*\*\*

#### Mental Fatigue Monitoring using Brain Dynamics Preferences

Yuanguang Pan, Avinash K Singh, Ivor W. Tsang, Chin-teng Lin

Driver's cognitive state of mental fatigue significantly affects driving performance and more importantly public safety. Previous studies leverage the response time (RT) as the metric for mental fatigue and aim at estimating the exact value of RT using electroencephalogram (EEG) signals within a regression model. However, due to the easily corrupted EEG signals and also non-smooth RTs during data collection, regular regression methods generally suffer from poor generalization performance. Considering that human response time is the reflection of brain dynamics preference rather than a single value, a novel model called Brain Dynamic Ranking (BDrank) has been proposed. BDrank could learn from brain dynamics preferences using EEG data robustly and preserve the ordering corresponding to RTs. BDrank model is based on the regularized alternative ordinal classification comparing to regular regression based practices. Furthermore, a transition matrix is introduced to characterize the reliability of each channel used in EEG data, which helps in learning brain dynamics preferences only from informative EEG channels.

nels. In order to handle large-scale EEG signals~and obtain higher generalization, an online-generalized Expectation Maximum (OnlineGEM) algorithm also has been proposed to update BDrank in an online fashion. Comprehensive empirical analysis on EEG signals from 44 participants shows that BDrank together with OnlineGEM achieves substantial improvements in reliability while simultaneously detecting possible less informative and noisy EEG channels.

\*\*\*\*\*

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

Wenda Zhou,Victor Veitch,Morgane Austern,Ryan P. Adams,Peter Orbanz

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

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.

\*\*\*\*\*

#### Conscious Inference for Object Detection

Jiahuan Zhou,Nikolaos Karianakis,Ying Wu,Gang Hua

Current Convolutional Neural Network (CNN)-based object detection models adopt strictly feedforward inference to predict the final detection results. However, the widely used one-way inference is agnostic to the global image context and the interplay between input image and task semantics. In this work, we present a general technique to improve off-the-shelf CNN-based object detection models in the inference stage without re-training, architecture modification or ground-truth requirements. We propose an iterative, bottom-up and top-down inference mechanism, which is named conscious inference, as it is inspired by prevalent models for human consciousness with top-down guidance and temporal persistence. While the downstream pass accumulates category-specific evidence over time, it subsequently affects the proposal calculation and the final detection. Feature activations

are updated in line with no additional memory cost. Our approach advances the state of the art using popular detection models (Faster-RCNN, YOLOv2, YOLOv3) on 2D object detection and 6D object pose estimation.

\*\*\*\*\*

#### Aggregated Momentum: Stability Through Passive Damping

James Lucas, Shengyang Sun, Richard Zemel, Roger Grosse

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. Large damping 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.

\*\*\*\*\*

#### Recovering the Lowest Layer of Deep Networks with High Threshold Activations

Surbhi Goel, Rina Panigrahy

Giving provable guarantees for learning neural networks is a core challenge of machine learning theory. Most prior work gives parameter recovery guarantees for one hidden layer networks, however, the networks used in practice have multiple non-linear layers. In this work, we show how we can strengthen such results to deeper networks -- we address the problem of uncovering the lowest layer in a deep neural network under the assumption that the lowest layer uses a high threshold before applying the activation, the upper network can be modeled as a well-behaved polynomial and the input distribution is gaussian.

\*\*\*\*\*

#### ON BREIMAN'S DILEMMA IN NEURAL NETWORKS: SUCCESS AND FAILURE OF NORMALIZED MARGINS

Yifei HUANG, Yuan YAO, Weizhi ZHU

A belief persists long in machine learning that enlargement of margins over training data accounts for the resistance of models to overfitting by increasing the robustness. Yet Breiman shows a dilemma (Breiman, 1999) that a uniform improvement on margin distribution \emph{does not} necessarily reduce generalization error. In this paper, we revisit Breiman's dilemma in deep neural networks with recently proposed normalized margins using Lipschitz constant bound by spectral norm products. With both simplified theory and extensive experiments, Breiman's dilemma is shown to rely on dynamics of normalized margin distributions, that reflects the trade-off between model expression power and data complexity. When the complexity of data is comparable to the model expression power in the sense that training and test data share similar phase transitions in normalized margin dynamics, two efficient ways are derived via classic margin-based generalization bounds to successfully predict the trend of generalization error. On the other hand, over-expressed models that exhibit uniform improvements on training normalized margins may lose such a prediction power and fail to prevent the overfitting.

\*\*\*\*\*

#### Activity Regularization for Continual Learning

Quang H. Pham, Steven C. H. Hoi

While deep neural networks have achieved remarkable successes, they suffer the well-known catastrophic forgetting issue when switching from existing tasks to tackle a new one. In this paper, we study continual learning with deep neural networks that learn from tasks arriving sequentially. We first propose an approximated multi-task learning framework that unifies a family of popular regularization based continual learning methods. We then analyze the weakness of existing approaches, and propose a novel regularization method named "Activity Regularization



" (AR), which alleviates forgetting meanwhile keeping model's plasticity to acquire new knowledge. Extensive experiments show that our method outperform state-of-the-art methods and effectively overcomes catastrophic forgetting.

\*\*\*\*\*

#### A Guider Network for Multi-Dual Learning

Wenpeng Hu,Zhengwei Tao,Zhanxing Zhu,Bing Liu,Zhou Lin,Jinwen Ma,Dongyan Zhao,Rui Yan

A large amount of parallel data is needed to train a strong neural machine translation (NMT) system. This is a major challenge for low-resource languages. Building on recent work on unsupervised and semi-supervised methods, we propose a multi-dual learning framework to improve the performance of NMT by using an almost infinite amount of available monolingual data and some parallel data of other languages. Since our framework involves multiple languages and components, we further propose a timing optimization method that uses reinforcement learning (RL) to optimally schedule the different components in order to avoid imbalanced training. Experimental results demonstrate the validity of our model, and confirm its superiority to existing dual learning methods.

\*\*\*\*\*

#### Calibration of neural network logit vectors to combat adversarial attacks

Oliver Goldstein

Adversarial examples remain an issue for contemporary neural networks. This paper draws on Background Check (Perello-Nieto et al., 2016), a technique in model calibration, to assist two-class neural networks in detecting adversarial examples, using the one dimensional difference between logit values as the underlying measure. This method interestingly tends to achieve the highest average recall on image sets that are generated with large perturbation vectors, which is unlike the existing literature on adversarial attacks (Cubuk et al., 2017). The proposed method does not need knowledge of the attack parameters or methods at training time, unlike a great deal of the literature that uses deep learning based methods to detect adversarial examples, such as Metzen et al. (2017), imbuing the proposed method with additional flexibility.

\*\*\*\*\*

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

Daewoo Kim,Sangwoo Moon,David Hostallero,Wan Ju Kang,Taeyoung Lee,Kyunghwan Son,Yung Yi

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

Sequence to sequence (seq2seq) models have become a popular framework for neural sequence prediction. While traditional seq2seq models are trained by Maximum Li

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.

\*\*\*\*\*

#### Probabilistic Semantic Embedding

Yue Jiao, Jonathon Hare, Adam Prügél-Bennett

We present an extension of a variational auto-encoder that creates semantically rich coupled probabilistic latent representations that capture the semantics of multiple modalities of data. We demonstrate this model through experiments using images and textual descriptors as inputs and images as outputs. Our latent representations are not only capable of driving a decoder to generate novel data, but can also be used directly for annotation or classification. Using the MNIST and Fashion-MNIST datasets we show that the embedding not only provides better reconstruction and classification performance than the current state-of-the-art, but it also allows us to exploit the semantic content of the pretrained word embedding spaces to do tasks such as image generation from labels outside of those seen during training.

\*\*\*\*\*

#### 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

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

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.

\*\*\*\*\*

#### Compound Density Networks

Agustinus Kristiadi, Asja Fischer

Despite the huge success of deep neural networks (NNs), finding good mechanisms for quantifying their prediction uncertainty is still an open problem. It was recently shown, that using an ensemble of NNs trained with a proper scoring rule leads to results competitive to those of Bayesian NNs. This ensemble method can be

is understood as finite mixture model with uniform mixing weights. We build on this mixture model approach and increase its flexibility by replacing the fixed mixing weights by an adaptive, input-dependent distribution (specifying the probability of each component) represented by an NN, and by considering uncountably many mixture components. The resulting model can be seen as the continuous counterpart to mixture density networks and is therefore referred to as compound density network. We empirically show that the proposed model results in better uncertainty estimates and is more robust to adversarial examples than previous approaches.

\*\*\*\*\*

Learning from Positive and Unlabeled Data with a Selection Bias

Masahiro Kato, Takeshi Teshima, Junya Honda

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

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

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

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.

\*\*\*\*\*

#### Adversarial Sampling for Active Learning

Christoph Mayer,Radu Timofte

This paper proposes ASAL, a new pool based active learning method that generates high entropy samples. Instead of directly annotating the synthetic samples, ASAL searches similar samples from the pool and includes them for training. Hence, the quality of new samples is high and annotations are reliable. ASAL is particularly suitable for large data sets because it achieves a better run-time complexity (sub-linear) for sample selection than traditional uncertainty sampling (linear). We present a comprehensive set of experiments on two data sets and show that ASAL outperforms similar methods and clearly exceeds the established baseline (random sampling). In the discussion section we analyze in which situations ASAL performs best and why it is sometimes hard to outperform random sample selection. To the best of our knowledge this is the first adversarial active learning technique that is applied for multiple class problems using deep convolutional classifiers and demonstrates superior performance than random sample selection.

\*\*\*\*\*

#### On-Policy Trust Region Policy Optimisation with Replay Buffers

Dmitry Kangin,Nicolas Pugeault

Building upon the recent success of deep reinforcement learning methods, we investigate the possibility of on-policy reinforcement learning improvement by reusing the data from several consecutive policies. On-policy methods bring many benefits, such as ability to evaluate each resulting policy. However, they usually discard all the information about the policies which existed before. In this work, we propose adaptation of the replay buffer concept, borrowed from the off-policy learning setting, to the on-policy algorithms. To achieve this, the proposed algorithm generalises the Q-, value and advantage functions for data from multiple policies. The method uses trust region optimisation, while avoiding some of the common problems of the algorithms such as TRPO or ACKTR: it uses hyperparameters to replace the trust region selection heuristics, as well as the trainable covariance matrix instead of the fixed one. In many cases, the method not only improves the results comparing to the state-of-the-art trust region on-policy learning algorithms such as ACKTR and TRPO, but also with respect to their off-policy counterpart DDPG.

\*\*\*\*\*

#### Step-wise Sensitivity Analysis: Identifying Partially Distributed Representations for Interpretable Deep Learning

Botty Dimanov,Mateja Jamnik

In this paper, we introduce a novel method, called step-wise sensitivity analysis, which makes three contributions towards increasing the interpretability of Deep Neural Networks (DNNs). First, we are the first to suggest a methodology that aggregates results across input stimuli to gain model-centric results. Second, we linearly approximate the neuron activation and propose to use the outlier weights to identify distributed code. Third, our method constructs a dependency graph of the relevant neurons across the network to gain fine-grained understanding of the nature and interactions of DNN's internal features. The dependency graph illustrates shared subgraphs that generalise across 10 classes and can be clustered into semantically related groups. This is the first step towards building decision trees as an interpretation of learned representations.

\*\*\*\*\*

#### Neural Networks for Modeling Source Code Edits

Rui Zhao,David Bieber,Kevin Swersky,Daniel Tarlow

Programming languages are emerging as a challenging and interesting domain for machine learning. A core task, which has received significant attention in recent years, is building generative models of source code. However, to our knowledge, previous generative models have always been framed in terms of generating static snapshots of code. In this work, we instead treat source code as a dynamic object and tackle the problem of modeling the edits that software developers make to source code files. This requires extracting intent from previous edits and leveraging it to generate subsequent edits. We develop several neural networks and use synthetic data to test their ability to learn challenging edit patterns that require strong generalization. We then collect and train our models on a large-scale dataset consisting of millions of fine-grained edits from thousands of Python developers.

\*\*\*\*\*

#### ChainGAN: A sequential approach to GANs

Safwan Hossain,Kiarash Jamali,Yuchen Li,Frank Rudzicz

We propose a new architecture and training methodology for generative adversarial networks. Current approaches attempt to learn the transformation from a noise sample to a generated data sample in one shot. Our proposed generator architecture, called ChainGAN, uses a two-step process. It first attempts to transform a noise vector into a crude sample, similar to a traditional generator. Next, a chain of networks, called editors, attempt to sequentially enhance this sample. We train each of these units independently, instead of with end-to-end backpropagation on the entire chain. Our model is robust, efficient, and flexible as we can apply it to various network architectures. We provide rationale for our choices and experimentally evaluate our model, achieving competitive results on several datasets.

\*\*\*\*\*

#### Phrase-Based Attentions

Phi Xuan Nguyen,Shafiq Joty

Most state-of-the-art neural machine translation systems, despite being different in architectural skeletons (e.g., recurrence, convolutional), share an indispensable feature: the Attention. However, most existing attention methods are token-based and ignore the importance of phrasal alignments, the key ingredient for the success of phrase-based statistical machine translation. In this paper, we propose novel phrase-based attention methods to model n-grams of tokens as attention entities. We incorporate our phrase-based attentions into the recently proposed Transformer network, and demonstrate that our approach yields improvements of 1.3 BLEU for English-to-German and 0.5 BLEU for German-to-English translation tasks, and 1.75 and 1.35 BLEU points in English-to-Russian and Russian-to-English translation tasks on WMT newstest2014 using WMT'16 training data.

\*\*\*\*\*

#### ON THE USE OF CONVOLUTIONAL AUTO-ENCODER FOR INCREMENTAL CLASSIFIER LEARNING IN CONTEXT AWARE ADVERTISEMENT

Tin Lay Nwe,Shudong Xie,Balaji Nataraj,Yiqun Li,Joo-Hwee Lim

Context Aware Advertisement (CAA) is a type of advertisement appearing on websites or mobile apps. The advertisement is targeted on specific group of users and/or the content displayed on the websites or apps. This paper focuses on classifying images displayed on the websites by incremental learning classifier with Deep Convolutional Neural Network (DCNN) especially for Context Aware Advertisement (CAA) framework. Incrementally learning new knowledge with DCNN leads to catastrophic forgetting as previously stored

information is replaced with new information. To prevent catastrophic forgetting, part of previously learned knowledge should be stored for the life time of incremental classifier. Storing information for life time involves privacy and legal concerns especially in context aware advertising framework. Here, we propose an incremental classifier learning method which addresses privacy and legal concerns while taking care of catastrophic forgetting problem. We conduct experiments on different datasets including CIFAR-100. Experimental results show that proposed system achieves relatively high performance compared to the state-of-the-art incremental learning methods.

\*\*\*\*\*

#### Sinkhorn AutoEncoders

Giorgio Patrini, Marcello Carioni, Patrick Forré, Samarth Bhargav, Max Welling, Riann van den Berg, Tim Genewein, Frank Nielsen

Optimal Transport offers an alternative to maximum likelihood for learning generative autoencoding models. We show how this principle dictates the minimization of the Wasserstein distance between the encoder aggregated posterior and the prior, plus a reconstruction error. We prove that in the non-parametric limit the autoencoder generates the data distribution if and only if the two distributions match exactly, and that the optimum can be obtained by deterministic autoencoders.

We then introduce the Sinkhorn AutoEncoder (SAE), which casts the problem into Optimal Transport on the latent space. The resulting Wasserstein distance is minimized by backpropagating through the Sinkhorn algorithm.

SAE models the aggregated posterior as an implicit distribution and therefore does not need a reparameterization trick for gradients estimation. Moreover, it requires virtually no adaptation to different prior distributions. We demonstrate its flexibility by considering models with hyperspherical and Dirichlet priors, as well as a simple case of probabilistic programming. SAE matches or outperforms other autoencoding models in visual quality and FID scores.

\*\*\*\*\*

#### Equi-normalization of Neural Networks

Pierre Stock, Benjamin Graham, Rémi Gribonval, Hervé Jégou

Modern neural networks are over-parametrized. In particular, each rectified linear hidden unit can be modified by a multiplicative factor by adjusting input and output weights, without changing the rest of the network. Inspired by the Sinkhorn-Knopp algorithm, we introduce a fast iterative method for minimizing the  $\ell_2$  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.

\*\*\*\*\*

#### Investigating CNNs' Learning Representation under label noise

Ryuichiro Hataya, Hideki Nakayama

Deep convolutional neural networks (CNNs) are known to be robust against label noise on extensive datasets. However, at the same time, CNNs are capable of memorizing all labels even if they are random, which means they can memorize corrupted labels. Are CNNs robust or fragile to label noise? Much of researches focusing on such memorization uses class-independent label noise to simulate label corruption, but this setting is simple and unrealistic. In this paper, we investigate the behavior of CNNs under class-dependently simulated label noise, which is generated based on the conceptual distance between classes of a large dataset (i.e., ImageNet-1k). Contrary to previous knowledge, we reveal CNNs are more robust to such class-dependent label noise than class-independent label noise. We also demonstrate the networks under class-dependent noise situations learn similar representation to the no noise situation, compared to class-independent noise situations.

\*\*\*\*\*

Information Theoretic lower bounds on negative log likelihood

Luis A. Lastras-Montaña

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.

\*\*\*\*\*

Finding Mixed Nash Equilibria of Generative Adversarial Networks

Ya-Ping Hsieh, Chen Liu, Volkan Cevher

We reconsider the training objective of Generative Adversarial Networks (GANs) from the mixed Nash Equilibria (NE) perspective. Inspired by the classical prox methods, we develop a novel algorithmic framework for GANs via an infinite-dimensional two-player game and prove rigorous convergence rates to the mixed NE. We then propose a principled procedure to reduce our novel prox methods to simple sampling routines, leading to practically efficient algorithms. Finally, we provide experimental evidence that our approach outperforms methods that seek pure strategy equilibria, such as SGD, Adam, and RMSProp, both in speed and quality.

\*\*\*\*\*

Universal Attacks on Equivariant Networks

Amit Deshpande, Sandesh Kamath, K V Subrahmanyam

Adversarial attacks on neural networks perturb the input at test time in order to fool trained and deployed neural network models. Most attacks such as gradient-based Fast Gradient Sign Method (FGSM) by Goodfellow et al. 2015 and DeepFool by Moosavi-Dezfooli et al. 2016 are input-dependent, small, pixel-wise perturbations, and they give different attack directions for different inputs. On the other hand, universal adversarial attacks are input-agnostic and the same attack works for most inputs. Translation or rotation-equivariant neural network models provide one approach to prevent universal attacks based on simple geometric transformations. In this paper, we observe an interesting spectral property shared by all of the above input-dependent, pixel-wise adversarial attacks on translation and rotation-equivariant networks. We exploit this property to get a single universal attack direction that fools the model on most inputs. Moreover, we show how to compute this universal attack direction using principal components of the existing input-dependent attacks on a very small sample of test inputs. We complement our empirical results by a theoretical justification, using matrix concentration inequalities and spectral perturbation bounds. We also empirically observe that the top few principal adversarial attack directions are nearly orthogonal to the top few principal invariant directions.

\*\*\*\*\*

Adversarial Attacks on Node Embeddings

Aleksandar Bojchevski, Stephan Günnemann

The goal of network representation learning is to learn low-dimensional node embeddings that capture the graph structure and are useful for solving downstream tasks. However, despite the proliferation of such methods there is currently no study of their robustness to adversarial attacks. We provide the first adversarial vulnerability analysis on the widely used family of methods based on random wa

lks. We derive efficient adversarial perturbations that poison the network structure and have a negative effect on both the quality of the embeddings and the downstream tasks. We further show that our attacks are transferable since they generalize to many models, and are successful even when the attacker is restricted.

\*\*\*\*\*

#### Consistent Jumpy Predictions for Videos and Scenes

Ananya Kumar, S. M. Ali Eslami, Danilo Rezende, Marta Garnelo, Fabio Viola, Edward Lockhart, Murray Shanahan

Stochastic video prediction models take in a sequence of image frames, and generate a sequence of consecutive future image frames. These models typically generate future frames in an autoregressive fashion, which is slow and requires the input and output frames to be consecutive. We introduce a model that overcomes these drawbacks by generating a latent representation from an arbitrary set of frames that can then be used to simultaneously and efficiently sample temporally consistent frames at arbitrary time-points. For example, our model can "jump" and directly sample frames at the end of the video, without sampling intermediate frames. Synthetic video evaluations confirm substantial gains in speed and functionality without loss in fidelity. We also apply our framework to a 3D scene reconstruction dataset. Here, our model is conditioned on camera location and can sample consistent sets of images for what an occluded region of a 3D scene might look like, even if there are multiple possibilities for what that region might contain. Reconstructions and videos are available at <https://bit.ly/2O4Pc4R>.

\*\*\*\*\*

#### Layerwise Recurrent Autoencoder for General Real-world Traffic Flow Forecasting

Peize Zhao, Danfeng Cai, Shaokun Zhang, Feng Chen, Zhemin Zhang, Cheng Wang, Jonathan Li

Accurate spatio-temporal traffic forecasting is a fundamental task that has wide applications in city management, transportation area and financial domain. There are many factors that make this significant task also challenging, like: (1) maze-like road network makes the spatial dependency complex; (2) the traffic-time relationships bring non-linear temporal complication; (3) with the larger road network, the difficulty of flow forecasting grows. The prevalent and state-of-the-art methods have mainly been discussed on datasets covering relatively small districts and short time span, e.g., the dataset that is collected within a city during months. To forecast the traffic flow across a wide area and overcome the mentioned challenges, we design and propose a promising forecasting model called Layerwise Recurrent Autoencoder (LRA), in which a three-layer stacked autoencoder (SAE) architecture is used to obtain temporal traffic correlations and a recurrent neural networks (RNNs) model for prediction. The convolutional neural networks (CNNs) model is also employed to extract spatial traffic information within the transport topology for more accurate prediction. To the best of our knowledge, there is no general and effective method for traffic flow prediction in large area which covers a group of cities. The experiment is completed on such large scale real-world traffic datasets to show superiority. And a smaller dataset is exploited to prove universality of the proposed model. And evaluations show that our model outperforms the state-of-the-art baselines by 6% - 15%.

\*\*\*\*\*

#### Neural Random Projections for Language Modelling

Davide Nunes, Luis Antunes

Neural network-based language models deal with data sparsity problems by mapping the large discrete space of words into a smaller continuous space of real-valued vectors. By learning distributed vector representations for words, each training sample informs the neural network model about a combinatorial number of other patterns. In this paper, we exploit the sparsity in natural language even further by encoding each unique input word using a fixed sparse random representation.

These sparse codes are then projected onto a smaller embedding space which allows for the encoding of word occurrences from a possibly unknown vocabulary, along with the creation of more compact language models using a reduced number of parameters.



ameters. We investigate the properties of our encoding mechanism empirically, by evaluating its performance on the widely used Penn Treebank corpus. We show that guaranteeing approximately equidistant vector representations for unique discrete inputs is enough to provide the neural network model with enough information to learn --and make use-- of distributed representations for these inputs.

\*\*\*\*\*

#### Neural Speed Reading with Structural-Jump-LSTM

Christian Hansen, Casper Hansen, Stephen Alstrup, Jakob Grue Simonsen, Christina Lioma

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 De von Hjelm

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

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.

\*\*\*\*\*

The effectiveness of layer-by-layer training using the information bottleneck principle

Adar Elad, Doron Haviv, Yochai Blau, Tomer Michaeli

The recently proposed information bottleneck (IB) theory of deep nets suggests that during training, each layer attempts to maximize its mutual information (MI) with the target labels (so as to allow good prediction accuracy), while minimizing its MI with the input (leading to effective compression and thus good generalization). To date, evidence of this phenomenon has been indirect and aroused controversy due to theoretical and practical complications. In particular, it has been pointed out that the MI with the input is theoretically infinite in many cases of interest, and that the MI with the target is fundamentally difficult to estimate in high dimensions. As a consequence, the validity of this theory has been questioned. In this paper, we overcome these obstacles by two means. First, as previously suggested, we replace the MI with the input by a noise-regularized version, which ensures it is finite. As we show, this modified penalty in fact acts as a form of weight decay regularization. Second, to obtain accurate (noise regularized) MI estimates between an intermediate representation and the input, we incorporate the strong prior-knowledge we have about their relation, into the recently proposed MI estimator of Belghazi et al. (2018). With this scheme, we are able to stably train each layer independently to explicitly optimize the IB functional. Surprisingly, this leads to enhanced prediction accuracy, thus directly validating the IB theory of deep nets for the first time.

\*\*\*\*\*

Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers

Yonatan Geifman, Guy Uziel, Ran El-Yaniv

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

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

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

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.

\*\*\*\*\*

#### Unseen Action Recognition with Unpaired Adversarial Multimodal Learning

AJ Piergiovanni, Michael S. Ryoo

In this paper, we present a method to learn a joint multimodal representation space that allows for the recognition of unseen activities in videos. We compare the effect of placing various constraints on the embedding space using paired text and video data. Additionally, we propose a method to improve the joint embedding space using an adversarial formulation with unpaired text and video data. In addition to testing on publicly available datasets, we introduce a new, large-scale text/video dataset. We experimentally confirm that learning such shared embedding space benefits three difficult tasks (i) zero-shot activity classification, (ii) unsupervised activity discovery, and (iii) unseen activity captioning.

\*\*\*\*\*

#### Connecting the Dots Between MLE and RL for Sequence Generation

Bowen Tan\*, Zhiting Hu\*, Zichao Yang, Ruslan Salakhutdinov, Eric P. Xing

Sequence generation models such as recurrent networks can be trained with a diverse set of learning algorithms. For example, maximum likelihood learning is simple and efficient, yet suffers from the exposure bias problem. Reinforcement learning like policy gradient addresses the problem but can have prohibitively poor exploration efficiency. A variety of other algorithms such as RAML, SPG, and data noising, have also been developed in different perspectives. This paper estab-

ishes a formal connection between these algorithms. We present a generalized entropy regularized policy optimization formulation, and show that the apparently divergent algorithms can all be reformulated as special instances of the framework, with the only difference being the configurations of reward function and a couple of hyperparameters. The unified interpretation offers a systematic view of the varying properties of exploration and learning efficiency. Besides, based on the framework, we present a new algorithm that dynamically interpolates among the existing algorithms for improved learning. Experiments on machine translation and text summarization demonstrate the superiority of the proposed algorithm.

\*\*\*\*\*

Temporal Gaussian Mixture Layer for Videos

AJ Piergiovanni, Michael S. Ryoo

We introduce a new convolutional layer named the Temporal Gaussian Mixture (TGM) layer and present how it can be used to efficiently capture longer-term temporal information in continuous activity videos. The TGM layer is a temporal convolutional layer governed by a much smaller set of parameters (e.g., location/variance of Gaussians) that are fully differentiable. We present our fully convolutional video models with multiple TGM layers for activity detection. The experiments on multiple datasets including Charades and MultiTHUMOS confirm the effectiveness of TGM layers, outperforming the state-of-the-arts.

\*\*\*\*\*

EXPLORATION OF EFFICIENT ON-DEVICE ACOUSTIC MODELING WITH NEURAL NETWORKS

Wonyong Sung, Lukas Lee, Jinwhan Park

Real-time speech recognition on mobile and embedded devices is an important application of neural networks. Acoustic modeling is the fundamental part of speech recognition and is usually implemented with long short-term memory (LSTM)-based recurrent neural networks (RNNs). However, the single thread execution of an LSTM RNN is extremely slow in most embedded devices because the algorithm needs to fetch a large number of parameters from the DRAM for computing each output sample. We explore a few acoustic modeling algorithms that can be executed very efficiently on embedded devices. These algorithms reduce the overhead of memory accesses using multi-timestep parallelization that computes multiple output samples at a time by reading the parameters only once from the DRAM. The algorithms considered are the quasi RNNs (QRNNs), Gated ConvNets, and diagonalized LSTMs. In addition, we explore neural networks that equip one-dimensional (1-D) convolution at each layer of these algorithms, and by which can obtain a very large performance increase in the QRNNs and Gated ConvNets. The experiments were conducted using two tasks, one is the connectionist temporal classification (CTC)-based end-to-end speech recognition on WSJ corpus and the other is the phoneme classification on TIMIT dataset. We not only significantly increase the execution speed but also obtain a much higher accuracy, compared to LSTM RNN-based modeling. Thus, this work can be applicable not only to embedded system-based implementations but also to server-based ones.

\*\*\*\*\*

What Would pi\* Do?: Imitation Learning via Off-Policy Reinforcement Learning

Siddharth Reddy, Anca D. Dragan, Sergey Levine

Learning to imitate expert actions given demonstrations containing image observations is a difficult problem in robotic control. The key challenge is generalizing behavior to out-of-distribution states that differ from those in the demonstrations. State-of-the-art imitation learning algorithms perform well in environments with low-dimensional observations, but typically involve adversarial optimization procedures, which can be difficult to use with high-dimensional image observations. We propose a remarkably simple alternative based on off-policy soft Q-learning, which we call soft Q imitation learning (SQIL, pronounced "skill"), that rewards the agent for matching demonstrated actions in demonstrated states. The key idea is initially filling the agent's experience replay buffer with demonstrations, where rewards are set to a positive constant, and setting rewards to zero in all additional experiences. We derive SQIL from first principles as a method for performing approximate inference under the MaxCausalEnt model of expert behavior. The approximate inference objective trades off between a pure behavior

ral cloning loss and a regularization term that incorporates information about state transitions via the soft Bellman error. Our experiments show that SQIL matches the state of the art in low-dimensional environments, and significantly outperforms prior work in playing video games from high-dimensional images.

\*\*\*\*\*

#### Auxiliary Variational MCMC

Raza Habib, David Barber

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>.

\*\*\*\*\*

#### Teacher Guided Architecture Search

Pouya Bashivan, Mark Tensen, James J DiCarlo

Strong improvements in neural network performance in vision tasks have resulted from the search of alternative network architectures, and prior work has shown that this search process can be automated and guided by evaluating candidate network performance following limited training ("Performance Guided Architecture Search" or PGAS). However, because of the large architecture search spaces and the high computational cost associated with evaluating each candidate model, further gains in computational efficiency are needed. Here we present a method termed Teacher Guided Search for Architectures by Generation and Evaluation (TG-SAGE) that produces up to an order of magnitude in search efficiency over PGAS methods. Specifically, TG-SAGE guides each step of the architecture search by evaluating the similarity of internal representations of the candidate networks with those of the (fixed) teacher network. We show that this procedure leads to significant reduction in required per-sample training and that, this advantage holds for two different search spaces of architectures, and two different search algorithms. We further show that in the space of convolutional cells for visual categorization, TG-SAGE finds a cell structure with similar performance as was previously found using other methods but at a total computational cost that is two orders of magnitude lower than Neural Architecture Search (NAS) and more than four times lower than progressive neural architecture search (PNAS). These results suggest that TG-SAGE can be used to accelerate network architecture search in cases where one has access to some or all of the internal representations of a teacher network of interest, such as the brain.

\*\*\*\*\*

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

James Jordon, Jinsung Yoon, Mihaela van der Schaar

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.

\*\*\*\*\*

### 3D-RelNet: Joint Object and Relational Network for 3D Prediction

Nilesh Kulkarni, Ishan Misra, Shubham Tulsiani, Abhinav Gupta

We propose an approach to predict the 3D shape and pose for the objects present in a scene. Existing learning based methods that pursue this goal make independent predictions per object, and do not leverage the relationships amongst them. We argue that reasoning about these relationships is crucial, and present an approach to incorporate these in a 3D prediction framework. In addition to independent per-object predictions, we predict pairwise relations in the form of relative 3D pose, and demonstrate that these can be easily incorporated to improve object level estimates. We report performance across different datasets (SUNCG, NYUv2), and show that our approach significantly improves over independent prediction approaches while also outperforming alternate implicit reasoning methods.

\*\*\*\*\*

### A fast quasi-Newton-type method for large-scale stochastic optimisation

Adrian Wills, Thomas B. Schön, Carl Jidling

During recent years there has been an increased interest in stochastic adaptations of limited memory quasi-Newton methods, which compared to pure gradient-based routines can improve the convergence by incorporating second order information. In this work we propose a direct least-squares approach conceptually similar to the limited memory quasi-Newton methods, but that computes the search direction in a slightly different way. This is achieved in a fast and numerically robust manner by maintaining a Cholesky factor of low dimension. This is combined with a stochastic line search relying upon fulfilment of the Wolfe condition in a backtracking manner, where the step length is adaptively modified with respect to the optimisation progress. We support our new algorithm by providing several theoretical results guaranteeing its performance. The performance is demonstrated on real-world benchmark problems which shows improved results in comparison with already established methods.

\*\*\*\*\*

### Total Style Transfer with a Single Feed-Forward Network

Minseong Kim, Hyun-Chul Choi

Recent image style transferring methods achieved arbitrary stylization with input content and style images. To transfer the style of an arbitrary image to a content image, these methods used a feed-forward network with a lowest-scaled feature transformer or a cascade of the networks with a feature transformer of a corresponding scale. However, their approaches did not consider either multi-scaled style in their single-scale feature transformer or dependency between the transformed feature statistics across the cascade networks. This shortcoming resulted in generating partially and inexact style transfer in the generated images. To overcome this limitation of partial style transfer, we propose a total style transferring method which transfers multi-scaled feature statistics through a single feed-forward process. First, our method transforms multi-scaled feature maps of a content image into those of a target style image by considering both inter-channel correlations in each single scaled feature map and inter-scale correlations between multi-scaled feature maps. Second, each transformed feature map is inserted into the decoder layer of the corresponding scale using skip-connection. Finally, the skip-connected multi-scaled feature maps are decoded into a stylized image through our trained decoder network.

\*\*\*\*\*

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

Marton Havasi, Robert Peharz, José Miguel Hernández-Lobato

While deep neural networks are a highly successful model class, their large memory footprint puts considerable strain on energy consumption, communication bandwidth

width, 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.

\*\*\*\*\*

#### Efficient Codebook and Factorization for Second Order Representation Learning

Pierre Jacob, David Picard, Aymeric Histace, Edouard Klein

Learning rich and compact representations is an open topic in many fields such as word embedding, visual question-answering, object recognition or image retrieval. Although deep neural networks (convolutional or not) have made a major breakthrough during the last few years by providing hierarchical, semantic and abstract representations for all of these tasks, these representations are not necessarily as rich as needed nor as compact as expected. Models using higher order statistics, such as bilinear pooling, provide richer representations at the cost of higher dimensional features. Factorization schemes have been proposed but without being able to reach the original compactness of first order models, or at a heavy loss in performances. This paper addresses these two points by extending factorization schemes to codebook strategies, allowing compact representations with the same dimensionality as first order representations, but with second order performances. Moreover, we extend this framework with a joint codebook and factorization scheme, granting a reduction both in terms of parameters and computation cost. This formulation leads to state-of-the-art results and compact second-order models with few additional parameters and intermediate representations with a dimension similar to that of first-order statistics.

\*\*\*\*\*

#### Manifold Mixup: Learning Better Representations by Interpolating Hidden States

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Aaron Courville, Ioannis Mitliagkas, Yoshua Bengio

Deep networks often perform well on the data distribution on which they are trained, yet give incorrect (and often very confident) answers when evaluated on points from off of the training distribution. This is exemplified by the adversarial examples phenomenon but can also be seen in terms of model generalization and domain shift. Ideally, a model would assign lower confidence to points unlike those from the training distribution. We propose a regularizer which addresses this issue by training with interpolated hidden states and encouraging the classifier to be less confident at these points. Because the hidden states are learned, this has an important effect of encouraging the hidden states for a class to be concentrated in such a way so that interpolations within the same class or between two different classes do not intersect with the real data points from other classes. This has a major advantage in that it avoids the underfitting which can result from interpolating in the input space. We prove that the exact condition for this problem of underfitting to be avoided by Manifold Mixup is that the dimensionality of the hidden states exceeds the number of classes, which is often the case in practice. Additionally, this concentration can be seen as making the features in earlier layers more discriminative. We show that despite requ

iring no significant additional computation, Manifold Mixup achieves large improvements over strong baselines in supervised learning, robustness to single-step adversarial attacks, semi-supervised learning, and Negative Log-Likelihood on held out samples.

\*\*\*\*\*

#### GO Gradient for Expectation-Based Objectives

Yulai Cong, Miaoyun Zhao, Ke Bai, Lawrence Carin

Within many machine learning algorithms, a fundamental problem concerns efficient calculation of an unbiased gradient wrt parameters  $\gamma$  for expectation-based objectives  $\mathbb{E}_{q(\gamma)}[f(y)]$ . 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 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

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.

\*\*\*\*\*

#### Second-Order Adversarial Attack and Certifiable Robustness

Bai Li, Changyou Chen, Wenlin Wang, Lawrence Carin

Adversarial training has been recognized as a strong defense against adversarial attacks. In this paper, we propose a powerful second-order attack method that reduces the accuracy of the defense model by Madry et al. (2017). We demonstrate that adversarial training overfits to the choice of the norm in the sense that it is only robust to the attack used for adversarial training, thus suggesting it has not achieved universal robustness. The effectiveness of our attack method motivates an investigation of provable robustness of a defense model. To this end, we introduce a framework that allows one to obtain a certifiable lower bound on the prediction accuracy against adversarial examples. We conduct experiments to show the effectiveness of our attack method. At the same time, our defense model achieves significant improvements compared to previous works under our proposed attack.

\*\*\*\*\*

#### BEHAVIOR MODULE IN NEURAL NETWORKS

Andrey Sakryukin, Yongkang Wong, Mohan S. Kankanhalli

Prefrontal cortex (PFC) is a part of the brain which is responsible for behavior repertoire. Inspired by PFC functionality and connectivity, as well as human behavior formation process, we propose a novel modular architecture of neural networks with a Behavioral Module (BM) and corresponding end-to-end training strategy. This approach allows the efficient learning of behaviors and preferences re



presentation. This property is particularly useful for user modeling (as for dialog agents) and recommendation tasks, as allows learning personalized representations of different user states. In the experiment with video games playing, the results show that the proposed method allows separation of main task's objectives and behaviors between different BMs. The experiments also show network extensibility through independent learning of new behavior patterns. Moreover, we demonstrate a strategy for an efficient transfer of newly learned BMs to unseen tasks.

\*\*\*\*\*

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

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

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.

\*\*\*\*\*

DOMAIN ADAPTATION VIA DISTRIBUTION AND REPRESENTATION MATCHING: A CASE STUDY ON TRAINING DATA SELECTION VIA REINFORCEMENT LEARNING

Miaofeng Liu, Yan Song, Hongbin Zou, Tong Zhang

Supervised models suffer from domain shifting where distribution mismatch across domains greatly affect model performance. Particularly, noise scattered in each domain has played a crucial role in representing such distribution, especially in various natural language processing (NLP) tasks. In addressing this issue, training data selection (TDS) has been proven to be a prospective way to train supervised models with higher performance and efficiency. Following the TDS methodology, in this paper, we propose a general data selection framework with representation learning and distribution matching simultaneously for domain adaptation on neural models. In doing so, we formulate TDS as a novel selection process based on a learned distribution from the input data, which is produced by a trainable

e selection distribution generator (SDG) that is optimized by reinforcement learning (RL). Then, the model trained by the selected data not only predicts the target domain data in a specific task, but also provides input for the value function of the RL. Experiments are conducted on three typical NLP tasks, namely, part-of-speech tagging, dependency parsing, and sentiment analysis. Results demonstrate the validity and effectiveness of our approach.

\*\*\*\*\*

#### Few-shot Classification on Graphs with Structural Regularized GCNs

Shengzhong Zhang, Ziang Zhou, Zengfeng Huang, Zhongyu Wei

We consider the fundamental problem of semi-supervised node classification in attributed graphs with a focus on \emph{few-shot} learning. Here, we propose Structural Regularized Graph Convolutional Networks (SRGCN), novel neural network architectures extending the well-known GCN structures by stacking transposed convolutional layers for reconstruction of input features. We add a reconstruction error term in the loss function as a regularizer. Unlike standard regularization such as  $L_1$  or  $L_2$ , which controls the model complexity by including a penalty term depends solely on parameters, our regularization function is parameterized by a trainable neural network whose structure depends on the topology of the underlying graph. The new approach effectively addresses the shortcomings of previous graph convolution-based techniques for learning classifiers in the few-shot regime and significantly improves generalization performance over original GCNs when the number of labeled samples is insufficient. Experimental studies on three challenging benchmarks demonstrate that the proposed approach has matched state-of-the-art results and can improve classification accuracies by a notable margin when there are very few examples from each class.

\*\*\*\*\*

#### Greedy Attack and Gumbel Attack: Generating Adversarial Examples for Discrete Data

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, Michael I. Jordan

We present a probabilistic framework for studying adversarial attacks on discrete data. Based on this framework, we derive a perturbation-based method, Greedy Attack, and a scalable learning-based method, Gumbel Attack, that illustrate various tradeoffs in the design of attacks. We demonstrate the effectiveness of these methods using both quantitative metrics and human evaluation on various state-of-the-art models for text classification, including a word-based CNN, a character-based CNN and an LSTM. As an example of our results, we show that the accuracy of character-based convolutional networks drops to the level of random selection by modifying only five characters through Greedy Attack.

\*\*\*\*\*

#### Whitening and Coloring Batch Transform for GANs

Aliaksandr Siarohin, Enver Sangineto, Nicu Sebe

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

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 funda

mental 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.

\*\*\*\*\*

Hierarchical Attention: What Really Counts in Various NLP Tasks

Zehao Dou,Zhihua Zhang

Attention mechanisms in sequence to sequence models have shown great ability and wonderful performance in various natural language processing (NLP) tasks, such as sentence embedding, text generation, machine translation, machine reading comprehension, etc. Unfortunately, existing attention mechanisms only learn either high-level or low-level features. In this paper, we think that the lack of hierarchical mechanisms is a bottleneck in improving the performance of the attention mechanisms, and propose a novel Hierarchical Attention Mechanism (Ham) based on the weighted sum of different layers of a multi-level attention.

Ham achieves a state-of-the-art BLEU score of 0.26 on Chinese poem generation task and a nearly 6.5% averaged improvement compared with the existing machine reading comprehension models such as BIDAf and Match-LSTM. Furthermore, our experiments and theorems reveal that Ham has greater generalization and representation ability than existing attention mechanisms.

\*\*\*\*\*

Sample-efficient policy learning in multi-agent Reinforcement Learning via meta-learning

Jialian Li, Hang Su, Jun Zhu

To gain high rewards in multi-agent scenes, it is sometimes necessary to understand other agents and make corresponding optimal decisions. We can solve these tasks by first building models for other agents and then finding the optimal policy with these models. To get an accurate model, many observations are needed and this can be sample-inefficient. What's more, the learned model and policy can overfit to current agents and cannot generalize if the other agents are replaced by new agents. In many practical situations, each agent we face can be considered as a sample from a population with a fixed but unknown distribution. Thus we can treat the task against some specific agents as a task sampled from a task distribution. We apply meta-learning method to build models and learn policies. Therefore when new agents come, we can adapt to them efficiently. Experiments on grid games show that our method can quickly get high rewards.

\*\*\*\*\*

Learning agents with prioritization and parameter noise in continuous state and action space

Rajesh Devaraddi, G. Srinivasaraghavan

Reinforcement Learning (RL) problem can be solved in two different ways - the Value function-based approach and the policy optimization-based approach - to eventually arrive at an optimal policy for the given environment. One of the recent breakthroughs in reinforcement learning is the use of deep neural networks as function approximators to approximate the value function or q-function in a reinforcement learning scheme. This has led to results with agents automatically learning how to play games like alpha-go showing better-than-human performance. Deep Q-learning networks (DQN) and Deep Deterministic Policy Gradient (DDPG) are two such methods that have shown state-of-the-art results in recent times. Among the many variants of RL, an important class of problems is where the state and action spaces are continuous --- autonomous robots, autonomous vehicles, optimal control are all examples of such problems that can lend themselves naturally to reinforcement based algorithms, and have continuous state and action spaces. In

this paper, we adapt and combine approaches such as DQN and DDPG in novel ways to outperform the earlier results for continuous state and action space problems.

We believe these results are a valuable addition to the fast-growing body of results on Reinforcement Learning, more so for continuous state and action space problems.

\*\*\*\*\*

#### Sparse Dictionary Learning by Dynamical Neural Networks

Tsung-Han Lin, Ping Tak Peter Tang

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  $\ell_1$ -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

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.

\*\*\*\*\*

#### Guiding Physical Intuition with Neural Stethoscopes

Fabian Fuchs, Oliver Groth, Adam Kosior, Alex Bewley, Markus Wulfmeier, Andrea Vedaldi, Ingmar Posner

Model interpretability and systematic, targeted model adaptation present central challenges in deep learning. In the domain of intuitive physics, we study the task of visually predicting stability of block towers with the goal of understanding and influencing the model's reasoning. Our contributions are two-fold. Firstly, we introduce neural stethoscopes as a framework for quantifying the degree of importance of specific factors of influence in deep networks as well as for actively promoting and suppressing information as appropriate. In doing so, we unify concepts from multitask learning as well as training with auxiliary and adversarial losses. Secondly, we deploy the stethoscope framework to provide an in-depth analysis of a state-of-the-art deep neural network for stability prediction, specifically examining its physical reasoning. We show that the baseline model is susceptible to being misled by incorrect visual cues. This leads to a performance breakdown to the level of random guessing when training on scenarios where visual cues are inversely correlated with stability. Using stethoscopes to promote meaningful feature extraction increases performance from 51% to 90% prediction accuracy. Conversely, training on an easy dataset where visual cues are positively correlated with stability, the baseline model learns a bias leading to poor

performance on a harder dataset. Using an adversarial stethoscope, the network is successfully de-biased, leading to a performance increase from 66% to 88%.

\*\*\*\*\*

Deep Denoising: Rate-Optimal Recovery of Structured Signals with a Deep Prior  
Reinhard Heckel, Wen Huang, Paul Hand, Vladislav Voroninski

Deep neural networks provide state-of-the-art performance for image denoising, where the goal is to recover a near noise-free image from a noisy image.

The underlying principle is that neural networks trained on large datasets have empirically been shown to be able to generate natural images well from a low-dimensional latent representation of the image.

Given such a generator network, or prior, a noisy image can be denoised by finding the closest image in the range of the prior.

However, there is little theory to justify this success, let alone to predict the denoising performance as a function of the network's parameters.

In this paper we consider the problem of denoising an image from additive Gaussian noise, assuming the image is well described by a deep neural network with ReLU activations functions, mapping a  $k$ -dimensional latent space to an  $n$ -dimensional image.

We state and analyze a simple gradient-descent-like iterative algorithm that minimizes a non-convex loss function, and provably removes a fraction of  $(1 - O(k/n))$  of the noise energy.

We also demonstrate in numerical experiments that this denoising performance is, indeed, achieved by generative priors learned from data.

\*\*\*\*\*

Gradient Acceleration in Activation Functions

Sangchul Hahn, Heeyoul Choi

Dropout has been one of standard approaches to train deep neural networks, and it is known to regularize large models to avoid overfitting. The effect of dropout has been explained by avoiding co-adaptation.

In this paper, however, we propose a new explanation of why dropout works and propose a new technique to design better activation functions. First, we show that dropout can be explained as an optimization technique to push the input towards the saturation area of nonlinear activation function by accelerating gradient information flowing even in the saturation area in backpropagation. Based on this explanation, we propose a new technique for activation functions,  $\{\text{gradient acceleration in activation function (GAAF)}\}$ , that accelerates gradients to flow even in the saturation area. Then, input to the activation function can climb onto the saturation area which makes the network more robust because the model converges on a flat region.

Experiment results support our explanation of dropout and confirm that the proposed GAAF technique improves performances with expected properties.

\*\*\*\*\*

NUTS: Network for Unsupervised Telegraphic Summarization

Chanakya Malireddy, Tirth Maniar, Sajal Maheshwari, Manish Shrivastava

Extractive summarization methods operate by ranking and selecting the sentences which best encapsulate the theme of a given document. They do not fare well in domains like fictional narratives where there is no central theme and core information is not encapsulated by a small set of sentences. For the purpose of reducing the size of the document while conveying the idea expressed by each sentence, we need more sentence specific methods. Telegraphic summarization, which selects short segments across several sentences, is better suited for such domains. Telegraphic summarization captures the plot better by retaining shorter versions of each sentence while not really concerning itself with grammatically linking these segments. In this paper, we propose an unsupervised deep learning network (NUTS) to generate telegraphic summaries.

We use multiple encoder-decoder networks and learn to drop portions of the text that are inferable from the chosen segments. The model is agnostic to both sentence length and style. We demonstrate that the summaries produced by our model show significant quantitative and qualitative improvement over those produced by existing methods and baselines.

\*\*\*\*\*

## Understanding GANs via Generalization Analysis for Disconnected Support

Masaaki Imaizumi, Kenji Fukumizu

This paper provides theoretical analysis of generative adversarial networks (GANs) to explain its advantages over other standard methods of learning probability measures. GANs learn a probability through observations, using the objective function with a generator and a discriminator. While many empirical results indicate that GANs can generate realistic samples, the reason for such successful performance remains unelucidated. This paper focuses the situation where the target probability measure satisfies the disconnected support property, which means a separate support of a probability, and relates it with the advantage of GANs. It is theoretically shown that, unlike other popular models, GANs do not suffer from the decrease of generalization performance caused by the disconnected support property. We rigorously quantify the generalization performance of GANs of a given architecture, and compare it with the performance of the other models. Based on the theory, we also provide a guideline for selecting deep network architecture for GANs. We demonstrate some numerical examples which support our results.

\*\*\*\*\*

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

Carson Eisenach, Haichuan Yang, Ji Liu, Han Liu

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  $\mathbb{R}^d$  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.

\*\*\*\*\*

## On the Margin Theory of Feedforward Neural Networks

Colin Wei, Jason Lee, Qiang Liu, Tengyu Ma

Past works have shown that, somewhat surprisingly, over-parametrization can help generalization in neural networks. Towards explaining this phenomenon, we adopt a margin-based perspective. We establish: 1) for multi-layer feedforward relu networks, the global minimizer of a weakly-regularized cross-entropy loss has the maximum normalized margin among all networks, 2) as a result, increasing the over-parametrization improves the normalized margin and generalization error bounds for deep networks. In the case of two-layer networks, an infinite-width neural network enjoys the best generalization guarantees. The typical infinite feature methods are kernel methods; we compare the neural net margin with that of kernel methods and construct natural instances where kernel methods have much weaker generalization guarantees. We validate this gap between the two approaches empirically. Finally, this infinite-neuron viewpoint is also fruitful for analyzing optimization. We show that a perturbed gradient flow on infinite-size networks finds a global optimizer in polynomial time.

\*\*\*\*\*

## Mol-CycleGAN - a generative model for molecular optimization

Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Michał Warcho

Designing a molecule with desired properties is one of the biggest challenges in drug development, as it requires optimization of chemical compound structures with respect to many complex properties. To augment the compound design process we introduce Mol-CycleGAN -- a CycleGAN-based model that generates optimized compounds with a chemical scaffold of interest. Namely, given a molecule our model generates a structurally similar one with an optimized value of the considered property. We evaluate the performance of the model on selected optimization objectives related to structural properties (presence of halogen groups, number of aromatic rings) and to a physicochemical property (penalized logP). In the task of optimization of penalized logP of drug-like molecules our model significantly outperforms previous results.

\*\*\*\*\*

code2seq: Generating Sequences from Structured Representations of Code

Uri Alon, Shaked Brody, Omer Levy, Eran Yahav

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>.

\*\*\*\*\*

Overfitting Detection of Deep Neural Networks without a Hold Out Set

Konrad Groh

Overfitting is an ubiquitous problem in neural network training and usually mitigated using a holdout data set.

Here we challenge this rationale and investigate criteria for overfitting without using a holdout data set.

Specifically, we train a model for a fixed number of epochs multiple times with varying fractions of randomized labels and for a range of regularization strengths.

A properly trained model should not be able to attain an accuracy greater than the fraction of properly labeled data points. Otherwise the model overfits.

We introduce two criteria for detecting overfitting and one to detect underfitting. We analyze early stopping, the regularization factor, and network depth.

In safety critical applications we are interested in models and parameter settings which perform well and are not likely to overfit. The methods of this paper allow characterizing and identifying such models.

\*\*\*\*\*

Generative Models from the perspective of Continual Learning

Timothée Lesort, Hugo Caselles-Dupré, Michael Garcia-Ortiz, Jean-François Goudou, David Filliat

Which generative model is the most suitable for Continual Learning? This paper aims at evaluating and comparing generative models on disjoint sequential image generation tasks. We investigate how several models learn and forget, considering various strategies: rehearsal, regularization, generative replay and fine-tuning. We used two quantitative metrics to estimate the generation quality and memory ability. We experiment with sequential tasks on three commonly used benchmarks for Continual Learning (MNIST, Fashion MNIST and CIFAR10). We found that among all models, the original GAN performs best and among Continual Learning strategies, generative replay outperforms all other methods. Even if we found satisfactory combinations on MNIST and Fashion MNIST, training generative models sequentially

lly on CIFAR10 is particularly instable, and remains a challenge.

\*\*\*\*\*

ADef: an Iterative Algorithm to Construct Adversarial Deformations

Rima Alaifari, Giovanni S. Alberti, Tandri Gauksson

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.

\*\*\*\*\*

Identifying Bias in AI using Simulation

Daniel McDuff, Roger Cheng, Ashish Kapoor

Machine learned models exhibit bias, often because the datasets used to train them are biased. This presents a serious problem for the deployment of such technology, as the resulting models might perform poorly on populations that are minorities within the training set and ultimately present higher risks to them. We propose to use high-fidelity computer simulations to interrogate and diagnose biases within ML classifiers. We present a framework that leverages Bayesian parameter search to efficiently characterize the high dimensional feature space and more quickly identify weakness in performance. We apply our approach to an example domain, face detection, and show that it can be used to help identify demographic biases in commercial face application programming interfaces (APIs).

\*\*\*\*\*

Revisiting Reweighted Wake-Sleep

Tuan Anh Le, Adam R. Kosiorek, N. Siddharth, Yee Whye Teh, Frank Wood

Discrete latent-variable models, while applicable in a variety of settings, can often be difficult to learn. Sampling discrete latent variables can result in high-variance gradient estimators for two primary reasons: 1) branching on the samples within the model, and 2) the lack of a pathwise derivative for the samples. While current state-of-the-art methods employ control-variate schemes for the former and continuous-relaxation methods for the latter, their utility is limited by the complexities of implementing and training effective control-variate schemes and the necessity of evaluating (potentially exponentially) many branch paths in the model. Here, we revisit the Reweighted Wake Sleep (RWS; Bornschein and Bengio, 2015) algorithm, and through extensive evaluations, show that it circumvents both these issues, outperforming current state-of-the-art methods in learning discrete latent-variable models. Moreover, we observe that, unlike the Importance-weighted Autoencoder, RWS learns better models and inference networks with increasing numbers of particles, and that its benefits extend to continuous latent-variable models as well. Our results suggest that RWS is a competitive, often preferable, alternative for learning deep generative models.

\*\*\*\*\*

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

Chulhee Yun, Suvrit Sra, Ali Jadbabaie

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



r deep linear networks, which unifies other results on this topic.

\*\*\*\*\*

#### Multi-way Encoding for Robustness to Adversarial Attacks

Donghyun Kim, Sarah Adel Bargal, Jianming Zhang, Stan Sclaroff

Deep models are state-of-the-art for many computer vision tasks including image classification and object detection. However, it has been shown that deep models are vulnerable to adversarial examples. We highlight how one-hot encoding directly contributes to this vulnerability and propose breaking away from this widely-used, but highly-vulnerable mapping. We demonstrate that by leveraging a different output encoding, multi-way encoding, we can make models more robust. Our approach makes it more difficult for adversaries to find useful gradients for generating adversarial attacks. We present state-of-the-art robustness results for black-box, white-box attacks, and achieve higher clean accuracy on four benchmark datasets: MNIST, CIFAR-10, CIFAR-100, and SVHN when combined with adversarial training. The strength of our approach is also presented in the form of an attack for model watermarking, raising challenges in detecting stolen models.

\*\*\*\*\*

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

Daniel McDuff, Ashish Kapoor

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

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.

\*\*\*\*\*

#### Riemannian Stochastic Gradient Descent for Tensor-Train Recurrent Neural Networks

Jun Qi, Chin-Hui Lee, Javier Tejedor

The Tensor-Train factorization (TTF) is an efficient way to compress large weight matrices of fully-connected layers and recurrent layers in recurrent neural networks (RNNs). However, high Tensor-Train ranks for all the core tensors of parameters need to be element-wise fixed, which results in an unnecessary redundancy of model parameters. This work applies Riemannian stochastic gradient descent (RSGD) to train core tensors of parameters in the Riemannian Manifold before finding vectors of lower Tensor-Train ranks for parameters. The paper first presents the RSGD algorithm with a convergence analysis and then tests it on more advanced

ed Tensor-Train RNNs such as bi-directional GRU/LSTM and Encoder-Decoder RNNs with a Tensor-Train attention model. The experiments on digit recognition and machine translation tasks suggest the effectiveness of the RSGD algorithm for Tensor-Train RNNs.

\*\*\*\*\*

#### The Singular Values of Convolutional Layers

Hanie Sedghi, Vineet Gupta, Philip M. Long

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

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

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.

\*\*\*\*\*

#### Learning Internal Dense But External Sparse Structures of Deep Neural Network

Yiqun Duan

Recent years have witnessed two seemingly opposite developments of deep convolutional neural networks (CNNs). On one hand, increasing the density of CNNs by adding cross-layer connections achieve higher accuracy. On the other hand, creating sparsity structures through regularization and pruning methods enjoys lower computational costs. In this paper, we bridge these two by proposing a new network structure with locally dense yet externally sparse connections. This new structure uses dense modules, as basic building blocks and then sparsely connects these modules via a novel algorithm during the training process. Experimental results demonstrate that the locally dense yet externally sparse structure could acquire competitive performance on benchmark tasks (CIFAR10, CIFAR100, and ImageNet) while keeping the network structure slim.

\*\*\*\*\*

#### FAST OBJECT LOCALIZATION VIA SENSITIVITY ANALYSIS

Mohammad K. Ebrahimpour, David C. Noelle

Deep Convolutional Neural Networks (CNNs) have been repeatedly shown to perform well on image classification tasks, successfully recognizing a broad array of objects when given sufficient training data. Methods for object localization, howe

ver, are still in need of substantial improvement. Common approaches to this problem involve the use of a sliding window, sometimes at multiple scales, providing input to a deep CNN trained to classify the contents of the window. In general, these approaches are time consuming, requiring many classification calculations. In this paper, we offer a fundamentally different approach to the localization of recognized objects in images. Our method is predicated on the idea that a deep CNN capable of recognizing an object must implicitly contain knowledge about object location in its connection weights. We provide a simple method to interpret classifier weights in the context of individual classified images. This method involves the calculation of the derivative of network generated activation patterns, such as the activation of output class label units, with regard to each input pixel, performing a sensitivity analysis that identifies the pixels that, in a local sense, have the greatest influence on internal representations and object recognition. These derivatives can be efficiently computed using a single backward pass through the deep CNN classifier, producing a sensitivity map of the image. We demonstrate that a simple linear mapping can be learned from sensitivity maps to bounding box coordinates, localizing the recognized object. Our experimental results, using real-world data sets for which ground truth localization information is known, reveal competitive accuracy from our fast technique.

\*\*\*\*\*

Downsampling leads to Image Memorization in Convolutional Autoencoders

Adityanarayanan Radhakrishnan, Caroline Uhler, Mikhail Belkin

Memorization of data in deep neural networks has become a subject of significant research interest.

In this paper, we link memorization of images in deep convolutional autoencoders to downsampling through strided convolution. To analyze this mechanism in a simpler setting, we train linear convolutional autoencoders and show that linear combinations of training data are stored as eigenvectors in the linear operator corresponding to the network when downsampling is used. On the other hand, networks without downsampling do not memorize training data. We provide further evidence that the same effect happens in nonlinear networks. Moreover, downsampling in nonlinear networks causes the model to not only memorize just linear combinations of images, but individual training images. Since convolutional autoencoder components are building blocks of deep convolutional networks, we envision that our findings will shed light on the important phenomenon of memorization in over-parameterized deep networks.

\*\*\*\*\*

Towards Understanding Regularization in Batch Normalization

Ping Luo, Xinjiang Wang, Wenqi Shao, Zhanglin Peng

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

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.

\*\*\*\*\*

Deep Generative Models for learning Coherent Latent Representations from Multi-Modal Data

Timo Korthals, Marc Hesse, Jürgen Leitner

The application of multi-modal generative models by means of a Variational Auto Encoder (VAE) is an upcoming research topic for sensor fusion and bi-directional modality exchange.

This contribution gives insights into the learned joint latent representation and shows that expressiveness and coherence are decisive properties for multi-modal datasets.

Furthermore, we propose a multi-modal VAE derived from the full joint marginal log-likelihood that is able to learn the most meaningful representation for ambiguous observations.

Since the properties of multi-modal sensor setups are essential for our approach but hardly available, we also propose a technique to generate correlated datasets from uni-modal ones.

\*\*\*\*\*

Attention, Learn to Solve Routing Problems!

Wouter Kool, Herke van Hoof, Max Welling

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

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

ts 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 constraints arising from learning dynamics and information processing.

\*\*\*\*\*

#### Meta-Learning with Individualized Feature Space for Few-Shot Classification

Chunrui Han, Shiguang Shan, Meina Kan, Shuzhe Wu, Xilin Chen

Meta-learning provides a promising learning framework to address few-shot classification tasks. In existing meta-learning methods, the meta-learner is designed to learn about model optimization, parameter initialization, or similarity metric. Differently, in this paper, we propose to learn how to create an individualized feature embedding specific to a given query image for better classifying, i.e., given a query image, a specific feature embedding tailored for its characteristics is created accordingly, leading to an individualized feature space in which the query image can be more accurately classified. Specifically, we introduce a kernel generator as meta-learner to learn to construct feature embedding for query images. The kernel generator acquires meta-knowledge of generating adequate convolutional kernels for different query images during training, which can generalize to unseen categories without fine-tuning. In two standard few-shot classification data sets, i.e. Omniglot, and \emph{mini}ImageNet, our method shows highly competitive performance.

\*\*\*\*\*

#### Meta-Learning Probabilistic Inference for Prediction

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, Richard Turner

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.

\*\*\*\*\*

#### Zero-Resource Multilingual Model Transfer: Learning What to Share

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, Claire Cardie

Modern natural language processing and understanding applications have enjoyed a great boost utilizing neural networks models. However, this is not the case for most languages especially low-resource ones with insufficient annotated training data. Cross-lingual transfer learning methods improve the performance on a low-resource target language by leveraging labeled data from other (source) languages, typically with the help of cross-lingual resources such as parallel corpora.

In this work, we propose a zero-resource multilingual transfer learning model that can utilize training data in multiple source languages, while not requiring target language training data nor cross-lingual supervision. Unlike most existing methods that only rely on language-invariant features for cross-lingual transfer, our approach utilizes both language-invariant and language-specific features in a coherent way. Our model leverages adversarial networks to learn language-invariant features and mixture-of-experts models to dynamically exploit the relation between the target language and each individual source language. This enables our model to learn effectively what to share between various languages in the

multilingual setup. It results in significant performance gains over prior art, as shown in an extensive set of experiments over multiple text classification and sequence tagging tasks including a large-scale real-world industry dataset.

\*\*\*\*\*

#### Ergodic Measure Preserving Flows

Yichuan Zhang, José Miguel Hernández-Lobato, Zoubin Ghahramani

Training probabilistic models with neural network components is intractable in most cases and requires to use approximations such as Markov chain Monte Carlo (MCMC), which is not scalable and requires significant hyper-parameter tuning, or mean-field variational inference (VI), which is biased. While there has been attempts at combining both approaches, the resulting methods have some important limitations in theory and in practice. As an alternative, we propose a novel method which is scalable, like mean-field VI, and, due to its theoretical foundation in ergodic theory, is also asymptotically accurate, like MCMC. We test our method on popular benchmark problems with deep generative models and Bayesian neural networks. Our results show that we can outperform existing approximate inference methods.

\*\*\*\*\*

#### Trellis Networks for Sequence Modeling

Shaojie Bai, J. Zico Kolter, Vladlen Koltun

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

Generative models for source 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.

\*\*\*\*\*

#### Scaling shared model governance via model splitting

Miljan Martić, Jan Leike, Andrew Trask, Matteo Hessel, Shane Legg, Pushmeet Kohli

Currently the only techniques for sharing governance of a deep learning model are homomorphic encryption and secure multiparty computation. Unfortunately, neither of these techniques is applicable to the training of large neural networks due to their large computational and communication overheads. As a scalable technique for shared model governance, we propose splitting deep learning model between multiple parties. This paper empirically investigates the security guarantee of this technique, which is introduced as the problem of model completion: Given the entire training data set or an environment simulator, and a subset of the parameters of a trained deep learning model, how much training is required to recover the model's original performance? We define a metric for evaluating the hardness of the model completion problem and study it empirically in both supervised learning on ImageNet and reinforcement learning on Atari and DeepMind Lab. Our experiments show that (1) the model completion problem is harder in reinforcement

ent learning than in supervised learning because of the unavailability of the trained agent's trajectories, and (2) its hardness depends not primarily on the number of parameters of the missing part, but more so on their type and location.

Our results suggest that model splitting might be a feasible technique for shared model governance in some settings where training is very expensive.

\*\*\*\*\*

PA-GAN: Improving GAN Training by Progressive Augmentation

Dan Zhang, Anna Khoreva

Despite recent progress, Generative Adversarial Networks (GANs) still suffer from training instability, requiring careful consideration of architecture design choices and hyper-parameter tuning. The reason for this fragile training behaviour is partially due to the discriminator performing well very quickly; its loss converges to zero, providing no reliable backpropagation signal to the generator.

In this work we introduce a new technique - progressive augmentation of GANs (PA-GAN) - that helps to overcome this fundamental limitation and improve the overall stability of GAN training. The key idea is to gradually increase the task difficulty of the discriminator by progressively augmenting its input space, thus enabling continuous learning of the generator. We show that the proposed progressive augmentation preserves the original GAN objective, does not bias the optimality of the discriminator and encourages the healthy competition between the generator and discriminator, leading to a better-performing generator. We experimentally demonstrate the effectiveness of the proposed approach on multiple benchmarks (MNIST, Fashion-MNIST, CIFAR10, CELEBA) for the image generation task.

\*\*\*\*\*

Interactive Parallel Exploration for Reinforcement Learning in Continuous Action Spaces

Whiyoung Jung, Giseung Park, Youngchul Sung

In this paper, a new interactive parallel learning scheme is proposed to enhance the performance of off-policy continuous-action reinforcement learning. In the proposed interactive parallel learning scheme, multiple identical learners with their own value-functions and policies share a common experience replay buffer, and search a good policy in collaboration with the guidance of the best policy information. The information of the best policy is fused in a soft manner by constructing an augmented loss function for policy update to enlarge the overall search space by the multiple learners. The guidance by the previous best policy and the enlarged search space by the proposed interactive parallel learning scheme enable faster and better policy search in the policy parameter space. Working algorithms are constructed by applying the proposed interactive parallel learning scheme to several off-policy reinforcement learning algorithms such as the twin delayed deep deterministic (TD3) policy gradient algorithm and the soft actor-critic (SAC) algorithm, and numerical results show that the constructed IPE-enhanced algorithms outperform most of the current state-of-the-art reinforcement learning algorithms for continuous action control.

\*\*\*\*\*

Adaptive Neural Trees

Ryutaro Tanno, Kai Arulkumaran, Daniel C. Alexander, Antonio Criminisi, Aditya Nori  
Deep neural networks and decision trees operate on largely separate paradigms; typically, the former performs representation learning with pre-specified architectures, while the latter is characterised by learning hierarchies over pre-specified features with data-driven architectures. We unite the two via adaptive neural trees (ANTs), a model that incorporates representation learning into edges, routing functions and leaf nodes of a decision tree, along with a backpropagation-based training algorithm that adaptively grows the architecture from primitive modules (e.g., convolutional layers). ANTs allow increased interpretability via hierarchical clustering, e.g., learning meaningful class associations, such as separating natural vs. man-made objects. We demonstrate this on classification and regression tasks, achieving over 99% and 90% accuracy on the MNIST and CIFAR-10 datasets, and outperforming standard neural networks, random forests and gradient boosted trees on the SARCOS dataset. Furthermore, ANT optimisation naturally adapts the architecture to the size and complexity of the training data.

\*\*\*\*\*

## State-Regularized Recurrent Networks

Cheng Wang, Mathias Niepert

Recurrent networks are a widely used class of neural architectures. They have, however, two shortcomings. First, it is difficult to understand what exactly they learn. Second, they tend to work poorly on sequences requiring long-term memorization, despite having this capacity in principle. We aim to address both shortcomings with a class of recurrent networks that use a stochastic state transition mechanism between cell applications. This mechanism, which we term state-regularization, makes RNNs transition between a finite set of learnable states. We show that state-regularization (a) simplifies the extraction of finite state automata modeling an RNN's state transition dynamics, and (b) forces RNNs to operate more like automata with external memory and less like finite state machines.

\*\*\*\*\*

## Ada-Boundary: Accelerating the DNN Training via Adaptive Boundary Batch Selection

Hwanjun Song, Sundong Kim, Minseok Kim, Jae-Gil Lee

Neural networks can converge faster with help from a smarter batch selection strategy. In this regard, we propose Ada-Boundary, a novel adaptive-batch selection algorithm that constructs an effective mini-batch according to the learning progress of the model. Our key idea is to present confusing samples what the true label is. Thus, the samples near the current decision boundary are considered as the most effective to expedite convergence. Taking advantage of our design, Ada-Boundary maintains its dominance in various degrees of training difficulty. We demonstrate the advantage of Ada-Boundary by extensive experiments using two convolutional neural networks for three benchmark data sets. The experiment results show that Ada-Boundary improves the training time by up to 31.7% compared with the state-of-the-art strategy and by up to 33.5% compared with the baseline strategy.

\*\*\*\*\*

## Siamese Capsule Networks

James O' Neill

Capsule Networks have shown encouraging results on defacto benchmark computer vision datasets such as MNIST, CIFAR and smallNORB. Although, they are yet to be tested on tasks where (1) the entities detected inherently have more complex internal representations and (2) there are very few instances per class to learn from and (3) where point-wise classification is not suitable. Hence, this paper carries out experiments on face verification in both controlled and uncontrolled settings that together address these points. In doing so we introduce Siamese Capsule Networks, a new variant that can be used for pairwise learning tasks. The model is trained using contrastive loss with l2-normalized capsule encoded pose features. We find that Siamese Capsule Networks perform well against strong baselines on both pairwise learning datasets, yielding best results in the few-shot learning setting where image pairs in the test set contain unseen subjects.

\*\*\*\*\*

## Effective Path: Know the Unknowns of Neural Network

Yuxian Qiu, Jingwen Leng, Yuhao Zhu, Quan Chen, Chao Li, Minyi Guo

Despite their enormous success, there is still no solid understanding of deep neural network's working mechanism. As such, researchers have demonstrated DNNs are vulnerable to small input perturbation, i.e., adversarial attacks. This work proposes the effective path as a new approach to exploring DNNs' internal organization. The effective path is an ensemble of synapses and neurons, which is reconstructed from a trained DNN using our activation-based backward algorithm. The per-image effective path can be aggregated to the class-level effective path, through which we observe that adversarial images activate effective path different from normal images. We propose an effective path similarity-based method to detect adversarial images and demonstrate its high accuracy and broad applicability.

\*\*\*\*\*

## Learning Recurrent Binary/Ternary Weights



Arash Ardakani, Zhengyun Ji, Sean C. Smithson, Brett H. Meyer, Warren J. Gross

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.

\*\*\*\*\*

Adversarial Vulnerability of Neural Networks Increases with Input Dimension  
Carl-Johann Simon-Gabriel, Yann Ollivier, Léon Bottou, Bernhard Schölkopf, David Lopez-Paz

Over the past four years, neural networks have been proven vulnerable to adversarial images: targeted but imperceptible image perturbations lead to drastically different predictions. We show that adversarial vulnerability increases with the gradients of the training objective when viewed as a function of the inputs. For most current network architectures, we prove that the L1-norm of these gradients grows as the square root of the input size. These nets therefore become increasingly vulnerable with growing image size. Our proofs rely on the network's weight distribution at initialization, but extensive experiments confirm that our conclusions still hold after usual training.

\*\*\*\*\*

PRUNING WITH HINTS: AN EFFICIENT FRAMEWORK FOR MODEL ACCELERATION

Wei Gao, Yi Wei, Quanguan Li, Hongwei Qin, Wanli Ouyang, Junjie Yan

In this paper, we propose an efficient framework to accelerate convolutional neural networks. We utilize two types of acceleration methods: pruning and hints. Pruning can reduce model size by removing channels of layers. Hints can improve the performance of student model by transferring knowledge from teacher model. We demonstrate that pruning and hints are complementary to each other. On one hand, hints can benefit pruning by maintaining similar feature representations. On the other hand, the model pruned from teacher networks is a good initialization for student model, which increases the transferability between two networks. Our approach performs pruning stage and hints stage iteratively to further improve the

performance. Furthermore, we propose an algorithm to reconstruct the parameters of hints layer and make the pruned model more suitable for hints. Experiments were conducted on various tasks including classification and pose estimation. Results on CIFAR-10, ImageNet and COCO demonstrate the generalization and superiority of our framework.

\*\*\*\*\*

Graph Neural Networks with Generated Parameters for Relation Extraction

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-seng Chua and Maosong Sun

Recently, progress has been made towards improving relational reasoning in machine learning field. Among existing models, graph neural networks (GNNs) is one of the most effective approaches for multi-hop relational reasoning. In fact, multi-hop relational reasoning is indispensable in many natural language processing tasks such as relation extraction. In this paper, we propose to generate the parameters of graph neural networks (GP-GNNs) according to natural language sentences, which enables GNNs to process relational reasoning on unstructured text inputs.

ts. We verify GP-GNNs in relation extraction from text. Experimental results on a human-annotated dataset and two distantly supervised datasets show that our model achieves significant improvements compared to the baselines. We also perform a qualitative analysis to demonstrate that our model could discover more accurate relations by multi-hop relational reasoning.

\*\*\*\*\*

microGAN: Promoting Variety through Microbatch Discrimination

Goncalo Mordido,Haojin Yang,Christoph Meinel

We propose to tackle the mode collapse problem in generative adversarial networks (GANs) by using multiple discriminators and assigning a different portion of each minibatch, called microbatch, to each discriminator. We gradually change each discriminator's task from distinguishing between real and fake samples to discriminating samples coming from inside or outside its assigned microbatch by using a diversity parameter  $\alpha$ . The generator is then forced to promote variety in each minibatch to make the microbatch discrimination harder to achieve by each discriminator. Thus, all models in our framework benefit from having variety in the generated set to reduce their respective losses. We show evidence that our solution promotes sample diversity since early training stages on multiple datasets.

\*\*\*\*\*

PPO-CMA: Proximal Policy Optimization with Covariance Matrix Adaptation

Perttu Härmäläinen,Amin Babadi,Xiaoxiao Ma,Jaakko Lehtinen

Proximal Policy Optimization (PPO) is a highly popular model-free reinforcement learning (RL) approach. However, in continuous state and actions spaces and a Gaussian policy -- common in computer animation and robotics -- PPO is prone to getting stuck in local optima. In this paper, we observe a tendency of PPO to prematurely shrink the exploration variance, which naturally leads to slow progress.

Motivated by this, we borrow ideas from CMA-ES, a black-box optimization method designed for intelligent adaptive Gaussian exploration, to derive PPO-CMA, a novel proximal policy optimization approach that expands the exploration variance on objective function slopes and only shrinks the variance when close to the optimum. This is implemented by using separate neural networks for policy mean and variance and training the mean and variance in separate passes. Our experiments demonstrate a clear improvement over vanilla PPO in many difficult OpenAI Gym MuJoCo tasks.

\*\*\*\*\*

Unsupervised Document Representation using Partition Word-Vectors Averaging

Vivek Gupta,Ankit Kumar Saw,Partha Pratim Talukdar,Praneeth Netrapalli

Learning effective document-level representation is essential in many important NLP tasks such as document classification, summarization, etc. Recent research has shown that simple weighted averaging of word vectors is an effective way to represent sentences, often outperforming complicated seq2seq neural models in many tasks. While it is desirable to use the same method to represent documents as well, unfortunately, the effectiveness is lost when representing long documents involving multiple sentences. One reason for this degradation is due to the fact that a longer document is likely to contain words from many different themes (or topics), and hence creating a single vector while ignoring all the thematic structure is unlikely to yield an effective representation of the document. This problem is less acute in single sentences and other short text fragments where presence of a single theme/topic is most likely. To overcome this problem, in this paper we present PSIF, a partitioned word averaging model to represent long documents. P-SIF retains the simplicity of simple weighted word averaging while taking a document's thematic structure into account. In particular, P-SIF learns topic-specific vectors from a document and finally concatenates them all to represent the overall document. Through our experiments over multiple real-world datasets and tasks, we demonstrate PSIF's effectiveness compared to simple weighted averaging and many other state-of-the-art baselines. We also show that PSIF is particularly effective in representing long multi-sentence documents. We will release PSIF's embedding source code and data-sets for reproducing results.

\*\*\*\*\*

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

Xiaoshuai Zhang, Yiping Lu, Jiaying Liu, Bin Dong

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.

\*\*\*\*\*

Classifier-agnostic saliency map extraction

Konrad Zolna, Krzysztof J. Geras, Kyunghyun Cho

Extracting saliency maps, which indicate parts of the image important to classification, requires many tricks to achieve satisfactory performance when using classifier-dependent methods. Instead, we propose classifier-agnostic saliency map extraction, which finds all parts of the image that any classifier could use, not just one given in advance. We observe that the proposed approach extracts higher quality saliency maps and outperforms existing weakly-supervised localization techniques, setting the new state of the art result on the ImageNet dataset.

\*\*\*\*\*

Improving Composition of Sentence Embeddings through the Lens of Statistical Relational Learning

Damien Sileo, Tim Van de Cruys, Camille Pradel, Philippe Muller

Various NLP problems -- such as the prediction of sentence similarity, entailment, and discourse relations -- are all instances of the same general task: the modeling of semantic relations between a pair of textual elements. We call them textual relational problems. A popular model for textual relational problems is to embed sentences into fixed size vectors and use composition functions (e.g. difference or concatenation) of those vectors as features for the prediction. Meanwhile, composition of embeddings has been a main focus within the field of Statistical Relational Learning (SRL) whose goal is to predict relations between entities (typically from knowledge base triples). In this work, we show that textual relational models implicitly use compositions from baseline SRL models. We show that such compositions are not expressive enough for several tasks (e.g. natural language inference). We build on recent SRL models to address textual relational problems, showing that they are more expressive, and can alleviate issues from simpler compositions. The resulting models significantly improve the state of the art in both transferable sentence representation learning and relation prediction.

\*\*\*\*\*

PPD: Permutation Phase Defense Against Adversarial Examples in Deep Learning

Mehdi Jafarnia-Jahromi, Tasmin Chowdhury, Hsin-Tai Wu, Sayandeep Mukherjee

Deep neural networks have demonstrated cutting edge performance on various tasks including classification. However, it is well known that adversarially designed imperceptible perturbation of the input can mislead advanced classifiers. In this paper, Permutation Phase Defense (PPD), is proposed as a novel method to resist adversarial attacks. PPD combines random permutation of the image with phase component of its Fourier transform. The basic idea behind this approach is to turn adversarial defense problems analogously into symmetric cryptography, which relies solely on safekeeping of the keys for security. In PPD, safe keeping of the selected permutation ensures effectiveness against adversarial attacks. Testing PPD on MNIST and CIFAR-10 datasets yielded state-of-the-art robustness against the most powerful adversarial attacks currently available.

\*\*\*\*\*

DiffraNet: Automatic Classification of Serial Crystallography Diffraction Patterns

Artur Souza, Leonardo B. Oliveira, Sabine Hollatz, Matt Feldman, Kunle Olukotun, Jame

s M. Holton,Aina E. Cohen,Luigi Nardi

Serial crystallography is the field of science that studies the structure and properties of crystals via diffraction patterns. In this paper, we introduce a new serial crystallography dataset generated through the use of a simulator; the synthetic images are labeled and they are both scalable and accurate. The resulting synthetic dataset is called DiffraNet, and it is composed of 25,000 512x512 grayscale labeled images. We explore several computer vision approaches for classification on DiffraNet such as standard feature extraction algorithms associated with Random Forests and Support Vector Machines but also an end-to-end CNN topology dubbed DeepFreak tailored to work on this new dataset. All implementations are publicly available and have been fine-tuned using off-the-shelf AutoML optimization tools for a fair comparison. Our best model achieves 98.5% accuracy. We believe that the DiffraNet dataset and its classification methods will have in the long term a positive impact in accelerating discoveries in many disciplines, including chemistry, geology, biology, materials science, metallurgy, and physics.

\*\*\*\*\*

BIGSAGE: unsupervised inductive representation learning of graph via bi-attended sampling and global-biased aggregating

Xin Luo,Hankz Hankui Zhuo

Different kinds of representation learning techniques on graph have shown significant effect in downstream machine learning tasks. Recently, in order to inductively learn representations for graph structures that is unobservable during training, a general framework with sampling and aggregating (GraphSAGE) was proposed by Hamilton and Ying and had been proved more efficient than transductive methods on fields like transfer learning or evolving dataset. However, GraphSAGE is incapable of selective neighbor sampling and lack of memory of known nodes that've been trained. To address these problems, we present an unsupervised method that samples neighborhood information attended by co-occurring structures and optimizes a trainable global bias as a representation expectation for each node in the given graph. Experiments show that our approach outperforms the state-of-the-art inductive and unsupervised methods for representation learning on graphs.

\*\*\*\*\*

The Effectiveness of Pre-Trained Code Embeddings

Ben Trevett,Donald Reay,N. K. Taylor

Word embeddings are widely used in machine learning based natural language processing systems. It is common to use pre-trained word embeddings which provide benefits such as reduced training time and improved overall performance. There has been a recent interest in applying natural language processing techniques to programming languages. However, none of this recent work uses pre-trained embeddings on code tokens. Using extreme summarization as the downstream task, we show that using pre-trained embeddings on code tokens provides the same benefits as it does to natural languages, achieving: over 1.9x speedup, 5% improvement in test loss, 4% improvement in F1 scores, and resistance to over-fitting. We also show that the choice of language used for the embeddings does not have to match that of the task to achieve these benefits and that even embeddings pre-trained on human languages provide these benefits to programming languages.

\*\*\*\*\*

Text Embeddings for Retrieval from a Large Knowledge Base

Tolgahan Cakaloglu,Christian Szegedy,Xiaowei Xu

Text embedding representing natural language documents in a semantic vector space can be used for document retrieval using nearest neighbor lookup. In order to study the feasibility of neural models specialized for retrieval in a semantically meaningful way, we suggest the use of the Stanford Question Answering Dataset (SQuAD) in an open-domain question answering context, where the first task is to find paragraphs useful for answering a given question. First, we compare the quality of various text-embedding methods on the performance of retrieval and give an extensive empirical comparison on the performance of various non-augmented base embedding with, and without IDF weighting. Our main results are that by training deep residual neural models specifically for retrieval purposes can yield

significant gains when it is used to augment existing embeddings. We also establish that deeper models are superior to this task. The best base baseline embeddings augmented by our learned neural approach improves the top-1 recall of the system by 14% in terms of the question side, and by 8% in terms of the paragraph side.

\*\*\*\*\*

#### Learning Implicit Generative Models by Teaching Explicit Ones

Chao Du, Kun Xu, Chongxuan Li, Jun Zhu, Bo Zhang

Implicit generative models are difficult to train as no explicit probability density functions are defined. Generative adversarial nets (GANs) propose a minimax framework to train such models, which suffer from mode collapse in practice due to the nature of the JS-divergence. In contrast, we propose a learning by teaching (LBT) framework to learn implicit models, which intrinsically avoid the mode collapse problem because of using the KL-divergence. In LBT, an auxiliary explicit model is introduced to learn the distribution defined by the implicit model while the later one's goal is to teach the explicit model to match the data distribution. LBT is formulated as a bilevel optimization problem, whose optimum implies that we obtain the maximum likelihood estimation of the implicit model. We adopt an unrolling approach to solve the challenging learning problem. Experimental results demonstrate the effectiveness of our method.

\*\*\*\*\*

#### VECTORIZATION METHODS IN RECOMMENDER SYSTEM

Qiang Sun, Bin Wang, Zizhou Gu, Yanwei Fu

The most used recommendation method is collaborative filtering, and the key part of collaborative filtering is to compute the similarity. The similarity based on co-occurrence of similar event is easy to implement and can be applied to almost all the situation. So when the word2vec model reach the state-of-art at a lower computation cost in NLP. An correspond model in recommender system item2vec is proposed and reach state-of-art in recommender system. It is easy to see that the position of user and item is interchangeable when their count size gap is not too much, we proposed a user2vec model and show its performance. The similarity based on co-occurrence information suffers from cold start, we proposed a content based similarity model based on doc2vec which is another technology in NLP.

\*\*\*\*\*

#### Decoupling Gating from Linearity

Yonathan Fiat, Eran Malach, Shai Shalev-Shwartz

The gap between the empirical success of deep learning and the lack of strong theoretical guarantees calls for studying simpler models. By observing that a ReLU neuron is a product of a linear function with a gate (the latter determines whether the neuron is active or not), where both share a jointly trained weight vector, we propose to decouple the two. We introduce GaLU networks – networks in which each neuron is a product of a Linear Unit, defined by a weight vector which is being trained, with a Gate, defined by a different weight vector which is not being trained. Generally speaking, given a base model and a simpler version of it, the two parameters that determine the quality of the simpler version are whether its practical performance is close enough to the base model and whether it is easier to analyze it theoretically. We show that GaLU networks perform similarly to ReLU networks on standard datasets and we initiate a study of their theoretical properties, demonstrating that they are indeed easier to analyze. We believe that further research of GaLU networks may be fruitful for the development of a theory of deep learning.

\*\*\*\*\*

#### DVOLVER: Efficient Pareto-Optimal Neural Network Architecture Search

Guillaume Michel, Mohammed Amine Alaoui, Alice Lebois, Amal Feriani, Mehdi Felhi

Automatic search of neural network architectures is a standing research topic. In addition to the fact that it presents a faster alternative to hand-designed architectures, it can improve their efficiency and for instance generate Convolutional Neural Networks (CNN) adapted for mobile devices. In this paper, we present a multi-objective neural architecture search method to find a family of CNN models with the best accuracy and computational resources tradeoffs, in a search space

are inspired by the state-of-the-art findings in neural search. Our work, called Dvolver, evolves a population of architectures and iteratively improves an approximation of the optimal Pareto front. Applying Dvolver on the model accuracy and on the number of floating points operations as objective functions, we are able to find, in only 2.5 days, a set of competitive mobile models on ImageNet. Amongst these models one architecture has the same Top-1 accuracy on ImageNet as NASNet-A mobile with 8% less floating point operations and another one has a Top-1 accuracy of 75.28% on ImageNet exceeding by 0.28% the best MobileNetV2 model for the same computational resources.

\*\*\*\*\*

On the Statistical and Information Theoretical Characteristics of DNN Representations

Daeyoung Choi, Wonjong Rhee, Kyungeun Lee, Changho Shin

It has been common to argue or imply that a regularizer can be used to alter a statistical property of a hidden layer's representation and thus improve generalization or performance of deep networks. For instance, dropout has been known to improve performance by reducing co-adaptation, and representational sparsity has been argued as a good characteristic because many data-generation processes have only a small number of factors that are independent. In this work, we analytically and empirically investigate the popular characteristics of learned representations, including correlation, sparsity, dead unit, rank, and mutual information, and disprove many of the conventional wisdoms. We first show that infinitely many Identical Output Networks (IONs) can be constructed for any deep network with a linear layer, where any invertible affine transformation can be applied to alter the layer's representation characteristics. The existence of ION proves that the correlation characteristics of representation can be either low or high for a well-performing network. Extensions to ReLU layers are provided, too. Then, we consider sparsity, dead unit, and rank to show that only loose relationships exist among the three characteristics. It is shown that a higher sparsity or additional dead units do not imply a better or worse performance when the rank of representation is fixed. We also develop a rank regularizer and show that neither representation sparsity nor lower rank is helpful for improving performance even when the data-generation process has only a small number of independent factors. Mutual information  $I(\mathbf{z}_l; \mathbf{x})$  and  $I(\mathbf{z}_l; \mathbf{y})$  are investigated as well, and we show that regularizers can affect  $I(\mathbf{z}_l; \mathbf{x})$  and thus indirectly influence the performance. Finally, we explain how a rich set of regularizers can be used as a powerful tool for performance tuning.

\*\*\*\*\*

Generating Images from Sounds Using Multimodal Features and GANs

Jeonghyun Lyu, Takashi Shinozaki, Kaoru Amano

Although generative adversarial networks (GANs) have enabled us to convert images from one domain to another similar one, converting between different sensory modalities, such as images and sounds, has been difficult. This study aims to propose a network that reconstructs images from sounds. First, video data with both images and sounds are labeled with pre-trained classifiers. Second, image and sound features are extracted from the data using pre-trained classifiers. Third, multimodal layers are introduced to extract features that are common to both the images and sounds. These layers are trained to extract similar features regardless of the input modality, such as images only, sounds only, and both images and sounds. Once the multimodal layers have been trained, features are extracted from input sounds and converted into image features using a feature-to-feature GAN. Finally, the generated image features are used to reconstruct images. Experimental results show that this method can successfully convert from the sound domain into the image domain. When we applied a pre-trained classifier to both the generated and original images, 31.9% of the examples had at least one of their top 10 labels in common, suggesting reasonably good image generation. Our results suggest that common representations can be learned for different modalities, and that proposed method can be applied not only to sound-to-image conversion but also to other conversions, such as from images to sounds.

\*\*\*\*\*

## Faster Training by Selecting Samples Using Embeddings

Santiago Gonzalez, Joshua Landgraf, Risto Miikkulainen

Long training times have increasingly become a burden for researchers by slowing down the pace of innovation, with some models taking days or weeks to train. In this paper, a new, general technique is presented that aims to speed up the training process by using a thinned-down training dataset. By leveraging autoencoders and the unique properties of embedding spaces, we are able to filter training datasets to include only those samples that matter the most. Through evaluation on a standard CIFAR-10 image classification task, this technique is shown to be effective. With this technique, training times can be reduced with a minimal loss in accuracy. Conversely, given a fixed training time budget, the technique was shown to improve accuracy by over 50%. This technique is a practical tool for achieving better results with large datasets and limited computational budgets.

\*\*\*\*\*

## Learning From the Experience of Others: Approximate Empirical Bayes in Neural Networks

Han Zhao, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, Geoff Gordon

Learning deep neural networks could be understood as the combination of representation learning and learning halfspaces. While most previous work aims to diversify representation learning by data augmentations and regularizations, we explore the opposite direction through the lens of empirical Bayes method. Specifically, we propose a matrix-variate normal prior whose covariance matrix has a Kronecker product structure to capture the correlations in learning different neurons through backpropagation. The prior encourages neurons to learn from the experience of others, hence it provides an effective regularization when training large networks on small datasets. To optimize the model, we design an efficient block coordinate descent algorithm with analytic solutions. Empirically, we show that the proposed method helps the network converge to better local optima that also generalize better, and we verify the effectiveness of the approach on both multi-class classification and multitask regression problems with various network structures.

\*\*\*\*\*

## Learning a SAT Solver from Single-Bit Supervision

Daniel Selsam, Matthew Lamm, Benedikt Böhler, Percy Liang, Leonardo de Moura, David L. Dill

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.

\*\*\*\*\*

## A fully automated periodicity detection in time series

Tom Puech, Matthieu Boussard

This paper presents a method to autonomously find periodicities in a signal. It is based on the same idea of using Fourier Transform and autocorrelation function presented in Vlachos et al. 2005. While showing interesting results this method does not perform well on noisy signals or signals with multiple periodicities. Thus, our method adds several new extra steps (hints clustering, filtering and detrending) to fix these issues. Experimental results show that the proposed method outperforms the state of the art algorithms.

\*\*\*\*\*

## Policy Generalization In Capacity-Limited Reinforcement Learning

Rachel A. Lerch, Chris R. Sims

Motivated by the study of generalization in biological intelligence, we examine reinforcement learning (RL) in settings where there are information-theoretic constraints placed on the learner's ability to represent a behavioral policy. We f

first show that the problem of optimizing expected utility within capacity-limited learning agents maps naturally to the mathematical field of rate-distortion (RD) theory. Applying the RD framework to the RL setting, we develop a new online RL algorithm, Capacity-Limited Actor-Critic, that learns a policy that optimizes a tradeoff between utility maximization and information processing costs. Using this algorithm in a 2D gridworld environment, we demonstrate two novel empirical results. First, at high information rates (high channel capacity), the algorithm achieves faster learning and discovers better policies compared to the standard tabular actor-critic algorithm. Second, we demonstrate that agents with capacity-limited policy representations avoid 'overfitting' and exhibit superior transfer to modified environments, compared to policies learned by agents with unlimited information processing resources. Our work provides a principled framework for the development of computationally rational RL agents.

\*\*\*\*\*

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

Eirikur Agustsson, Alexander Sage, Radu Timofte, Luc Van Gool

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 TWO-STEP ADVERSARIAL DEFENSE FOR DEEP NEURAL NETWORKS

Ting-Jui Chang, Yukun He, Peng Li

In recent years, deep neural networks have demonstrated outstanding performance in many machine learning tasks. However, researchers have discovered that the state-of-the-art models are vulnerable to adversarial examples: legitimate examples added by small perturbations which are unnoticeable to human eyes. Adversarial training, which augments the training data with adversarial examples during the training process, is a well known defense to improve the robustness of the model against adversarial attacks. However, this robustness is only effective to the same attack method used for adversarial training. Madry et al. (2017) suggest that effectiveness of iterative multi-step adversarial attacks and particularly that projected gradient descent (PGD) may be considered the universal first order adversary and applying the adversarial training with PGD implies resistance against many other first order attacks. However, the computational cost of the adversarial training with PGD and other multi-step adversarial examples is much higher than that of the adversarial training with other simpler attack techniques. In this paper, we show how strong adversarial examples can be generated only at a cost similar to that of two runs of the fast gradient sign method (FGSM), allowing defense against adversarial attacks with a robustness level comparable to that of the adversarial training with multi-step adversarial examples. We empirically demonstrate the effectiveness of the proposed two-step defense approach against different attack methods and its improvements over existing defense strategies.

\*\*\*\*\*

Localized random projections challenge benchmarks for bio-plausible deep learning



Bernd Illing, Wulfram Gerstner, Johanni Brea

Similar to models of brain-like computation, artificial deep neural networks rely on distributed coding, parallel processing and plastic synaptic weights. Training deep neural networks with the error-backpropagation algorithm, however, is considered bio-implausible. An appealing alternative to training deep neural networks is to use one or a few hidden layers with fixed random weights or trained with an unsupervised, local learning rule and train a single readout layer with a supervised, local learning rule. We find that a network of leaky-integrate-and-fire neurons with fixed random, localized receptive fields in the hidden layer and spike timing dependent plasticity to train the readout layer achieves 98.1% test accuracy on MNIST, which is close to the optimal result achievable with error-backpropagation in non-convolutional networks of rate neurons with one hidden layer. To support the design choices of the spiking network, we systematically compare the classification performance of rate networks with a single hidden layer, where the weights of this layer are either random and fixed, trained with unsupervised Principal Component Analysis or Sparse Coding, or trained with the backpropagation algorithm. This comparison revealed, first, that unsupervised learning does not lead to better performance than fixed random projections for large hidden layers on digit classification (MNIST) and object recognition (CIFAR10); second, networks with random projections and localized receptive fields perform significantly better than networks with all-to-all connectivity and almost reach the performance of networks trained with the backpropagation algorithm. The performance of these simple random projection networks is comparable to most current models of bio-plausible deep learning and thus provides an interesting benchmark for future approaches.

\*\*\*\*\*

Likelihood-based Permutation Invariant Loss Function for Probability Distributions

Masataro Asai

We propose a permutation-invariant loss function designed for the neural network reconstructing a set of elements without considering the order within its vector representation. Unlike popular approaches for encoding and decoding a set, our work does not rely on a carefully engineered network topology nor by any additional sequential algorithm. The proposed method, Set Cross Entropy, has a natural information-theoretic interpretation and is related to the metrics defined for sets. We evaluate the proposed approach in two object reconstruction tasks and a rule learning task.

\*\*\*\*\*

Escaping Flat Areas via Function-Preserving Structural Network Modifications

Yannic Kilcher, Gary Bécigneul, Thomas Hofmann

Hierarchically embedding smaller networks in larger networks, e.g. by increasing the number of hidden units, has been studied since the 1990s. The main interest was in understanding possible redundancies in the parameterization, as well as in studying how such embeddings affect critical points. We take these results as a point of departure to devise a novel strategy for escaping from flat regions of the error surface and to address the slow-down of gradient-based methods experienced in plateaus of saddle points. The idea is to expand the dimensionality of a network in a way that guarantees the existence of new escape directions. We call this operation the opening of a tunnel. One may then continue with the larger network either temporarily, i.e. closing the tunnel later, or permanently, i

.e.~iteratively growing the network, whenever needed. We develop our method for fully-connected as well as convolutional layers. Moreover, we present a practical version of our algorithm that requires no network structure modification and can be deployed as plug-and-play into any current deep learning framework. Experimentally, our method shows significant speed-ups.

\*\*\*\*\*

#### Incremental Hierarchical Reinforcement Learning with Multitask LMDPs

Adam C Earle, Andrew M Saxe, Benjamin Rosman

Exploration is a well known challenge in Reinforcement Learning. One principled way of overcoming this challenge is to find a hierarchical abstraction of the base problem and explore at these higher levels, rather than in the space of primitives. However, discovering a deep abstraction autonomously remains a largely unsolved problem, with practitioners typically hand-crafting these hierarchical control architectures. Recent work with multitask linear Markov decision processes, allows for the autonomous discovery of deep hierarchical abstractions, but operates exclusively in the offline setting. By extending this work, we develop an agent that is capable of incrementally growing a hierarchical representation, and using its experience to date to improve exploration.

\*\*\*\*\*

#### Prior Networks for Detection of Adversarial Attacks

Andrey Malinin, Mark Gales

Adversarial examples are considered a serious issue for safety critical applications of AI, such as finance, autonomous vehicle control and medicinal applications. Though significant work has resulted in increased robustness of systems to these attacks, systems are still vulnerable to well-crafted attacks. To address this problem

several adversarial attack detection methods have been proposed. However, system can still be vulnerable to adversarial samples that are designed to specifically evade these detection methods. One recent detection scheme that has shown good performance is based on uncertainty estimates derived from Monte-Carlo dropout ensembles. Prior Networks, a new method of estimating predictive uncertainty, have been shown to outperform Monte-Carlo dropout on a range of tasks. One of the advantages of this approach is that the behaviour of a Prior Network can be explicitly tuned to, for example, predict high uncertainty in regions where there are no training data samples. In this work Prior Networks are applied to adversarial attack detection using measures of uncertainty in a similar fashion to Monte-Carlo Dropout. Detection based on measures of uncertainty derived from DNNs and Monte-Carlo dropout ensembles are used as a baseline. Prior Networks are shown to significantly out-perform these baseline approaches over a range of adversarial attacks in both detection of whitebox and blackbox configurations. Even when the adversarial attacks are constructed with full knowledge of the detection mechanism, it is shown to be highly challenging to successfully generate an adversarial sample.

\*\*\*\*\*

#### Efficient Lifelong Learning with A-GEM

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, Mohamed Elhoseiny

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 classific

ation tasks under consideration. Our experiments on several standard lifelong learning benchmarks demonstrate that A-GEM has the best trade-off between accuracy and efficiency

\*\*\*\*\*

#### Learning to Coordinate Multiple Reinforcement Learning Agents for Diverse Query Reformulation

Rodrigo Nogueira, Jannis Bulian, Massimiliano Ciaramita

We propose a method to efficiently learn diverse strategies in reinforcement learning for query reformulation in the tasks of document retrieval and question answering. In the proposed framework an agent consists of multiple specialized sub-agents and a meta-agent that learns to aggregate the answers from sub-agents to produce a final answer. Sub-agents are trained on disjoint partitions of the training data, while the meta-agent is trained on the full training set. Our method makes learning faster, because it is highly parallelizable, and has better generalization performance than strong baselines, such as an ensemble of agents trained on the full data. We show that the improved performance is due to the increased diversity of reformulation strategies.

\*\*\*\*\*

#### Contextualized Role Interaction for Neural Machine Translation

Dirk Weissenborn, Douwe Kiela, Jason Weston, Kyunghyun Cho

Word inputs tend to be represented as single continuous vectors in deep neural networks. It is left to the subsequent layers of the network to extract relevant aspects of a word's meaning based on the context in which it appears. In this paper, we investigate whether word representations can be improved by explicitly incorporating the idea of latent roles. That is, we propose a role interaction layer (RIL) that consists of context-dependent (latent) role assignments and role-specific transformations. We evaluate the RIL on machine translation using two language pairs (En-De and En-Fi) and three datasets of varying size. We find that the proposed mechanism improves translation quality over strong baselines with limited amounts of data, but that the improvement diminishes as the size of data grows, indicating that powerful neural MT systems are capable of implicitly modeling role-word interaction by themselves. Our qualitative analysis reveals that the RIL extracts meaningful context-dependent roles and that it allows us to inspect more deeply the internal mechanisms of state-of-the-art neural machine translation systems.

\*\*\*\*\*

#### Talk The Walk: Navigating Grids in New York City through Grounded Dialogue

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, Douwe Kiela

We introduce "Talk The Walk", the first large-scale dialogue dataset grounded in action and perception. The task involves two agents (a 'guide' and a 'tourist') that communicate via natural language in order to achieve a common goal: having the tourist navigate to a given target location. The task and dataset, which are described in detail, are challenging and their full solution is an open problem that we pose to the community. We (i) focus on the task of tourist localization and develop the novel Masked Attention for Spatial Convolutions (MASC) mechanism that allows for grounding tourist utterances into the guide's map, (ii) show it yields significant improvements for both emergent and natural language communication, and (iii) using this method, we establish non-trivial baselines on the full task.

\*\*\*\*\*

#### Exploration by Uncertainty in Reward Space

Wei-Yang Qu, Yang Yu, Tang-Jie Lv, Ying-Feng Chen, Chang-Jie Fan

Efficient exploration plays a key role in reinforcement learning tasks. Commonly used dithering strategies, such as  $\epsilon$ -greedy, try to explore the action-state space randomly; this can lead to large demand for samples. In this paper, We propose an exploration method based on the uncertainty in reward space. There are two policies in this approach, the exploration policy is used for exploratory sampling in the environment, then the benchmark policy try to update by the data prove n by the exploration policy. Benchmark policy is used to provide the uncertainty in reward space, e.g. td-error, which guides the exploration policy updating. W

e apply our method on two grid-world environments and four Atari games. Experiment results show that our method improves learning speed and have a better performance than baseline policies

\*\*\*\*\*

Offline Deep models calibration with bayesian neural networks

Juan Maroñas, Roberto Paredes, Daniel Ramos

We apply Bayesian Neural Networks to improve calibration of state-of-the-art deep

neural networks. We show that, even with the most basic amortized approximate posterior distribution, and fast fully connected neural network for the likelihood,

the Bayesian framework clearly outperforms other simple maximum likelihood based solutions that have recently shown very good performance, as temperature scaling. As an example, we reduce the Expected Calibration

Error (ECE) from 0.52 to 0.24 on CIFAR-10 and from 4.28 to 2.456 on CIFAR-100

on two Wide ResNet with 96.13% and 80.39% accuracy respectively, which are among the best results published for this task. We demonstrate our robustness and

performance with experiments on a wide set of state-of-the-art computer vision models. Moreover, our approach acts off-line, and thus can be applied to any probabilistic model regardless of the limitations that the model may present during

training. This makes it suitable to calibrate systems that make use of pre-trained

deep neural networks that are expensive to train for a specific task, or to directly

train a calibrated deep convolutional model with Monte Carlo Dropout approximations, among others. However,

our method is still complementary with any Bayesian Neural Network for further improvement.

\*\*\*\*\*

A Solution to China Competitive Poker Using Deep Learning

Zhenxing Liu, Maoyu Hu, Zhangfei Zhang

Recently, deep neural networks have achieved superhuman performance in various games such as Go, chess and Shogi. Compared to Go, China Competitive Poker, also known as Dou Dizhu, is a type of imperfect information game, including hidden information, randomness, multi-agent cooperation and competition. It has become widespread and is now a national game in China. We introduce an approach to play China Competitive Poker using Convolutional Neural Network (CNN) to predict actions. This network is trained by supervised learning from human game records. Without any search, the network already beats the best AI program by a large margin, and also beats the best human amateur players in duplicate mode.

\*\*\*\*\*

A Variational Autoencoder for Probabilistic Non-Negative Matrix Factorisation

Steven Squires, Adam Prugel-Bennett, Mahesan Niranjan

We introduce and demonstrate the variational autoencoder (VAE) for probabilistic non-negative matrix factorisation (PAE-NMF). We design a network which can perform non-negative matrix factorisation (NMF) and add in aspects of a VAE to make the coefficients of the latent space probabilistic. By restricting the weights in the final layer of the network to be non-negative and using the non-negative Weibull distribution we produce a probabilistic form of NMF which allows us to generate new data and find a probability distribution that effectively links the latent and input variables. We demonstrate the effectiveness of PAE-NMF on three heterogeneous datasets: images, financial time series and genomic.

\*\*\*\*\*

Graph Matching Networks for Learning the Similarity of Graph Structured Objects

Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, Pushmeet Kohli

This paper addresses the challenging problem of retrieval and matching of graph structured objects, and makes two key contributions. First, we demonstrate how Graph Neural Networks (GNN), which have emerged as an effective model for various

s supervised prediction problems defined on structured data, can be trained to produce embedding of graphs in vector spaces that enables efficient similarity reasoning. Second, we propose a novel Graph Matching Network model that, given a pair of graphs as input, computes a similarity score between them by jointly reasoning on the pair through a new cross-graph attention-based matching mechanism. We demonstrate the effectiveness of our models on different domains including the challenging problem of control-flow-graph based function similarity search that plays an important role in the detection of vulnerabilities in software systems. The experimental analysis demonstrates that our models are not only able to exploit structure in the context of similarity learning but they can also outperform domain-specific baseline systems that have been carefully hand-engineered for these problems.

\*\*\*\*\*

#### Meta-Learning For Stochastic Gradient MCMC

Wenbo Gong, Yingzhen Li, José Miguel Hernández-Lobato

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.

\*\*\*\*\*

#### Cautious Deep Learning

Yotam Hechtlinger, Barnabas Poczos, Larry Wasserman

Most classifiers operate by selecting the maximum of an estimate of the conditional distribution  $p(y|x)$  where  $x$  stands for the features of the instance to be classified and  $y$  denotes its label. This often results in a hubristic bias: overconfidence in the assignment of a definite label. Usually, the observations are concentrated on a small volume but the classifier provides definite predictions for the entire space. We propose constructing conformal prediction sets which contain a set of labels rather than a single label. These conformal prediction sets contain the true label with probability  $1-\alpha$ . Our construction is based on  $p(x|y)$  rather than  $p(y|x)$  which results in a classifier that is very cautious: it outputs the null set --- meaning "I don't know" --- when the object does not resemble the training examples. An important property of our approach is that classes can be added or removed without having to retrain the classifier. We demonstrate the performance on the ImageNet ILSVRC dataset and the CelebA and IMDB-Wiki facial datasets using high dimensional features obtained from state of the art convolutional neural networks.

\*\*\*\*\*

#### Augmented Cyclic Adversarial Learning for Low Resource Domain Adaptation

Ehsan Hosseini-Asl, Yingbo Zhou, Caiming Xiong, Richard Socher

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

at 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 outperform 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.

\*\*\*\*\*

On Accurate Evaluation of GANs for Language Generation

Stanislau Semeniuta, Aliaksei Severyn, Sylvain Gelly

Generative Adversarial Networks (GANs) are a promising approach to language generation. The latest works introducing novel GAN models for language generation use n-gram based metrics for evaluation and only report single scores of the best run. In this paper, we argue that this often misrepresents the true picture and does not tell the full story, as GAN models can be extremely sensitive to the random initialization and small deviations from the best hyperparameter choice. In particular, we demonstrate that the previously used BLEU score is not sensitive to semantic deterioration of generated texts and propose alternative metrics that better capture the quality and diversity of the generated samples. We also conduct a set of experiments comparing a number of GAN models for text with a conventional Language Model (LM) and find that none of the considered models performs convincingly better than the LM.

\*\*\*\*\*

Wasserstein proximal of GANs

Alex Tong Lin, Wuchen Li, Stanley Osher, Guido Montufar

We introduce a new method for training GANs by applying the Wasserstein-2 metric proximal on the generators.

The approach is based on the gradient operator induced by optimal transport, which connects the geometry of sample space and parameter space in implicit deep generative models. From this theory, we obtain an easy-to-implement regularizer for the parameter updates. Our experiments demonstrate that this method improves the speed and stability in training GANs in terms of wall-clock time and Fréchet Inception Distance (FID) learning curves.

\*\*\*\*\*

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

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 prop

osal. 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

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 1.9x 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.

\*\*\*\*\*

Spatial-Winograd Pruning Enabling Sparse Winograd Convolution

Jiecao Yu, Jongsoo Park, Maxim Naumov

Deep convolutional neural networks (CNNs) are deployed in various applications but demand immense computational requirements. Pruning techniques and Winograd convolution are two typical methods to reduce the CNN computation. However, they cannot be directly combined because Winograd transformation fills in the sparsity resulting from pruning. Li et al. (2017) propose sparse Winograd convolution in which weights are directly pruned in the Winograd domain, but this technique is not very practical because Winograd-domain retraining requires low learning rates and hence significantly longer training time. Besides, Liu et al. (2018) move the ReLU function into the Winograd domain, which can help increase the weight sparsity but requires changes in the network structure. To achieve a high Winograd-domain weight sparsity without changing network structures, we propose a new pruning method, spatial-Winograd pruning. As the first step, spatial-domain weights are pruned in a structured way, which efficiently transfers the spatial-domain sparsity into the Winograd domain and avoids Winograd-domain retraining. For the next step, we also perform pruning and retraining directly in the Winograd domain but propose to use an importance factor matrix to adjust weight importance and weight gradients. This adjustment makes it possible to effectively retrain the pruned Winograd-domain network without changing the network structure. For the three models on the datasets of CIFAR-10, CIFAR-100, and ImageNet, our proposed method can achieve the Winograd-domain sparsities of 63%, 50%, and 74%, respectively.

\*\*\*\*\*

Pixel Chem: A Representation for Predicting Material Properties with Neural Network

Shuqian Ye, Yanheng Xu, Jiechun Liang, Hao Xu, Shuhong Cai, Shixin Liu, Xi Zhu

In this work we developed a new representation of the chemical information for the machine learning models, with benefits from both the real space (R-space) and

energy space (K-space). Different from the previous symmetric matrix presentations, the charge transfer channel based on Pauling's electronegativity is derived from the dependence on real space distance and orbitals for the hetero atomic structures. This representation can work for the bulk materials as well as the low dimensional nano materials, and can map the R-space and K-space into the pixel space (P-space) by training and testing 130k structures. P-space can well reproduce the R-space quantities within error 0.53. This new asymmetric matrix representation double the information storage than the previous symmetric representations. This work provides a new dimension for the computational chemistry towards the machine learning architecture.

\*\*\*\*\*

Skip-gram word embeddings in hyperbolic space

Matthias Leimeister, Benjamin J. Wilson

Embeddings of tree-like graphs in hyperbolic space were recently shown to surpass their Euclidean counterparts in performance by a large margin.

Inspired by these results, we present an algorithm for learning word embeddings in hyperbolic space from free text. An objective function based on the hyperbolic distance is derived and included in the skip-gram negative-sampling architecture from word2vec. The hyperbolic word embeddings are then evaluated on word similarity and analogy benchmarks. The results demonstrate the potential of hyperbolic word embeddings, particularly in low dimensions, though without clear superiority over their Euclidean counterparts. We further discuss subtleties in the formulation of the analogy task in curved spaces.

\*\*\*\*\*

Psychophysical vs. learnt texture representations in novelty detection

Michael Grunwald, Matthias Hermann, Fabian Freiberg, Matthias O. Franz

Parametric texture models have been applied successfully to synthesize artificial images. Psychophysical studies show that under defined conditions observers are unable to differentiate between model-generated and original natural textures.

In industrial applications the reverse case is of interest: a texture analysis system should decide if human observers are able to discriminate between a reference and a novel texture. For example, in case of inspecting decorative surfaces the detection of visible texture anomalies without any prior knowledge is required. Here, we implemented a human-vision-inspired novelty detection approach. Assuming that the features used for texture synthesis are important for human texture perception, we compare psychophysical as well as learnt texture representations based on activations of a pretrained CNN in a novelty detection scenario. Additionally, we introduce a novel objective function to train one-class neural networks for novelty detection and compare the results to standard one-class SVM approaches. Our experiments clearly show the differences between human-vision-inspired texture representations and learnt features in detecting visual anomalies. Based on a digital print inspection scenario we show that psychophysical texture representations are able to outperform CNN-encoded features.

\*\*\*\*\*

Learning to encode spatial relations from natural language

Tiago Ramalho, Tomas Kocisky, Frederic Besse, S. M. Ali Eslami, Gabor Melis, Fabio Viola, Phil Blunsom, Karl Moritz Hermann

Natural language processing has made significant inroads into learning the semantics of words through distributional approaches, however representations learnt via these methods fail to capture certain kinds of information implicit in the real world. In particular, spatial relations are encoded in a way that is inconsistent with human spatial reasoning and lacking invariance to viewpoint changes. We present a system capable of capturing the semantics of spatial relations such as behind, left of, etc from natural language. Our key contributions are a novel multi-modal objective based on generating images of scenes from their textual descriptions, and a new dataset on which to train it. We demonstrate that internal representations are robust to meaning preserving transformations of descriptions (paraphrase invariance), while viewpoint invariance is an emergent property of the system.

\*\*\*\*\*



## Meta-Learning with Domain Adaptation for Few-Shot Learning under Domain Shift

Doyen Sahoo, Hung Le, Chenghao Liu, Steven C. H. Hoi

Few-Shot Learning (learning with limited labeled data) aims to overcome the limitations of traditional machine learning approaches which require thousands of labeled examples to train an effective model. Considered as a hallmark of human intelligence, the community has recently witnessed several contributions on this topic, in particular through meta-learning, where a model learns how to learn an effective model for few-shot learning. The main idea is to acquire prior knowledge from a set of training tasks, which is then used to perform (few-shot) test tasks. Most existing work assumes that both training and test tasks are drawn from the same distribution, and a large amount of labeled data is available in the training tasks. This is a very strong assumption which restricts the usage of meta-learning strategies in the real world where ample training tasks following the same distribution as test tasks may not be available. In this paper, we propose a novel meta-learning paradigm wherein a few-shot learning model is learnt, which simultaneously overcomes domain shift between the train and test tasks via a adversarial domain adaptation. We demonstrate the efficacy the proposed method through extensive experiments.

\*\*\*\*\*

## Pooling Is Neither Necessary nor Sufficient for Appropriate Deformation Stability in CNNs

Avraham Ruderman, Neil C. Rabinowitz, Ari S. Morcos, Daniel Zoran

Many of our core assumptions about how neural networks operate remain empirically untested. One common assumption is that convolutional neural networks need to be stable to small translations and deformations to solve image recognition tasks. For many years, this stability was baked into CNN architectures by incorporating interleaved pooling layers. Recently, however, interleaved pooling has largely been abandoned. This raises a number of questions: Are our intuitions about deformation stability right at all? Is it important? Is pooling necessary for deformation invariance? If not, how is deformation invariance achieved in its absence? In this work, we rigorously test these questions, and find that deformation stability in convolutional networks is more nuanced than it first appears: (1) Deformation invariance is not a binary property, but rather that different tasks require different degrees of deformation stability at different layers. (2) Deformation stability is not a fixed property of a network and is heavily adjusted over the course of training, largely through the smoothness of the convolutional filters. (3) Interleaved pooling layers are neither necessary nor sufficient for achieving the optimal form of deformation stability for natural image classification. (4) Pooling confers \emph{too much} deformation stability for image classification at initialization, and during training, networks have to learn to \emph{counteract} this inductive bias. Together, these findings provide new insights into the role of interleaved pooling and deformation invariance in CNNs, and demonstrate the importance of rigorous empirical testing of even our most basic assumptions about the working of neural networks.

\*\*\*\*\*

## Gaussian-gated LSTM: Improved convergence by reducing state updates

Matthew Thornton, Jithendar Anumula, Shih-Chii Liu

Recurrent neural networks can be difficult to train on long sequence data due to the well-known vanishing gradient problem. Some architectures incorporate methods to reduce RNN state updates, therefore allowing the network to preserve memory over long temporal intervals. To address these problems of convergence, this paper proposes a timing-gated LSTM RNN model, called the Gaussian-gated LSTM (g-LSTM). The time gate controls when a neuron can be updated during training, enabling longer memory persistence and better error-gradient flow. This model captures long-temporal dependencies better than an LSTM and the time gate parameters can be learned even from non-optimal initialization values. Because the time gate limits the updates of the neuron state, the number of computes needed for the network update is also reduced. By adding a computational budget term to the training loss, we can obtain a network which further reduces the number of computes by at least 10x. Finally, by employing a temporal curriculum learning schedule for

For the g-LSTM, we can reduce the convergence time of the equivalent LSTM network on long sequences.

\*\*\*\*\*

#### Better Generalization with On-the-fly Dataset Denoising

Jiaming Song, Tengyu Ma, Michael Auli, Yann Dauphin

Memorization in over-parameterized neural networks can severely hurt generalization in the presence of mislabeled examples. However, mislabeled examples are hard to avoid in extremely large datasets. We address this problem using the implicit regularization effect of stochastic gradient descent with large learning rates, which we find to be able to separate clean and mislabeled examples with remarkable success using loss statistics. We leverage this to identify and on-the-fly discard mislabeled examples using a threshold on their losses. This leads to On-the-fly Data Denoising (ODD), a simple yet effective algorithm that is robust to mislabeled examples, while introducing almost zero computational overhead. Empirical results demonstrate the effectiveness of ODD on several datasets containing artificial and real-world mislabeled examples.

\*\*\*\*\*

#### SnapQuant: A Probabilistic and Nested Parameterization for Binary Networks

Kuan Wang, Hao Zhao, Anbang Yao, Aojun Zhou, Dawei Sun, Yurong Chen

In this paper, we study the problem of training real binary weight networks (without layer-wise or filter-wise scaling factors) from scratch under the Bayesian deep learning perspective, meaning that the final objective is to approximate the posterior distribution of binary weights rather than reach a point estimation.

The proposed method, named as SnapQuant, has two intriguing features: (1) The posterior distribution is parameterized as a policy network trained with a reinforcement learning scheme. During the training phase, we generate binary weights on-the-fly since what we actually maintain is the policy network, and all the binary weights are used in a burn-after-reading style. At the testing phase, we can sample binary weight instances for a given recognition architecture from the learnt policy network. (2) The policy network, which has a nested parameter structure consisting of layer-wise, filter-wise and kernel-wise parameter sharing designs, is applicable to any neural network architecture. Such a nested parameterization explicitly and hierarchically models the joint posterior distribution of binary weights. The performance of SnapQuant is evaluated with several visual recognition tasks including ImageNet. The code will be made publicly available.

\*\*\*\*\*

#### Graph Wavelet Neural Network

Bingbing Xu, Huawei Shen, Qi Cao, Yunqi Qiu, Xueqi Cheng

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.

\*\*\*\*\*

#### Cramer-Wold AutoEncoder

Jacek Tabor, Szymon Knop, Przemysław Spurek, Igor Podolak, Marcin Mazur, Stanisław Jastrzebski

Assessing distance between the true and the sample distribution is a key component of many state of the art generative models, such as Wasserstein Autoencoder (WAE). Inspired by prior work on Sliced-Wasserstein Autoencoders (SWAE) and kernel smoothing we construct a new generative model - Cramer-Wold AutoEncoder (CWAE). CWAE cost function, based on introduced Cramer-Wold distance between samples, has a simple closed-form in the case of normal prior. As a consequence, while simplifying the optimization procedure (no need of sampling necessary to evaluate the distance function in the training loop), CWAE performance matches quant

itatively and qualitatively that of WAE-MMD (WAE using maximum mean discrepancy based distance function) and often improves upon SWAE.

\*\*\*\*\*

#### 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

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

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  $l_\infty$  norm  $\epsilon = 0.1$ : 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.

\*\*\*\*\*

#### DeepTwist: Learning Model Compression via Occasional Weight Distortion

Dongsoo Lee, Parichay Kapoor, Byeongwook Kim

Model compression has been introduced to reduce the required hardware resources while maintaining the model accuracy. Lots of techniques for model compression, such as pruning, quantization, and low-rank approximation, have been suggested along with different inference implementation characteristics. Adopting model compression is, however, still challenging because the design complexity of model compression is rapidly increasing due to additional hyper-parameters and computation overhead in order to achieve a high compression ratio. In this paper, we propose a simple and efficient model compression framework called DeepTwist which distorts weights in an occasional manner without modifying the underlying training algorithms. The ideas of designing weight distortion functions are intuitive and straightforward given formats of compressed weights. We show that our proposed framework improves compression rate significantly for pruning, quantization, and low-rank approximation techniques while the efforts of additional retraining and/or hyper-parameter search are highly reduced. Regularization effects of Deep

Twist are also reported.

\*\*\*\*\*

#### Countdown Regression: Sharp and Calibrated Survival Predictions

Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam Shah, Andrew Ng

Personalized probabilistic forecasts of time to event (such as mortality) can be crucial in decision making, especially in the clinical setting. Inspired by ideas from the meteorology literature, we approach this problem through the paradigm of maximizing sharpness of prediction distributions, subject to calibration. In regression problems, it has been shown that optimizing the continuous ranked probability score (CRPS) instead of maximum likelihood leads to sharper prediction distributions while maintaining calibration. We introduce the Survival-CRPS, a generalization of the CRPS to the time to event setting, and present right-censored and interval-censored variants. To holistically evaluate the quality of predicted distributions over time to event, we present the scale agnostic Survival-AUPRC evaluation metric, an analog to area under the precision-recall curve. We apply these ideas by building a recurrent neural network for mortality prediction, using an Electronic Health Record dataset covering millions of patients. We demonstrate significant benefits in models trained by the Survival-CRPS objective instead of maximum likelihood.

\*\*\*\*\*

#### Synthnet: Learning synthesizers end-to-end

Florin Schimbinschi, Christian Walder, Sarah Erfani, James Bailey

Learning synthesizers and generating music in the raw audio domain is a challenging task. We investigate the learned representations of convolutional autoregressive generative models. Consequently, we show that mappings between musical notes and the harmonic style (instrument timbre) can be learned based on the raw audio music recording and the musical score (in binary piano roll format). Our proposed architecture, SynthNet uses minimal training data (9 minutes), is substantially better in quality and converges 6 times faster than the baselines. The quality of the generated waveforms (generation accuracy) is sufficiently high that they are almost identical to the ground truth. Therefore, we are able to directly measure generation error during training, based on the RMSE of the Constant-Q transform. Mean opinion scores are also provided. We validate our work using 7 distinct harmonic styles and also provide visualizations and links to all generated audio.

\*\*\*\*\*

#### Are Generative Classifiers More Robust to Adversarial Attacks?

Yingzhen Li, John Bradshaw, Yash Sharma

There is a rising interest in studying the robustness of deep neural network classifiers against adversaries, with both advanced attack and defence techniques being actively developed. However, most recent work focuses on discriminative classifiers, which only model the conditional distribution of the labels given the inputs. In this paper, we propose and investigate the deep Bayes classifier, which improves classical naive Bayes with conditional deep generative models. We further develop detection methods for adversarial examples, which reject inputs with low likelihood under the generative model. Experimental results suggest that deep Bayes classifiers are more robust than deep discriminative classifiers, and that the proposed detection methods are effective against many recently proposed attacks.

\*\*\*\*\*

#### Training Variational Auto Encoders with Discrete Latent Representations using Importance Sampling

Alexander Bartler, Felix Wiewel, Bin Yang, Lukas Mauch

The Variational Auto Encoder (VAE) is a popular generative latent variable model that is often applied for representation learning.

Standard VAEs assume continuous valued

latent variables and are trained by maximization

of the evidence lower bound (ELBO). Conventional methods obtain a

differentiable estimate of the ELBO with reparametrized sampling and

optimize it with Stochastic Gradient Descend (SGD). However, this is not possible if we want to train VAEs with discrete valued latent variables, since reparametrized sampling is not possible. Till now, there exist no simple solutions to circumvent this problem. In this paper, we propose an easy method to train VAEs with binary or categorically valued latent representations. Therefore, we use a differentiable estimator for the ELBO which is based on importance sampling. In experiments, we verify the approach and train two different VAEs architectures with Bernoulli and Categorical distributed latent representations on two different benchmark datasets. ■

\*\*\*\*\*

Towards the first adversarially robust neural network model on MNIST  
Lukas Schott, Jonas Rauber, Matthias Bethge, Wieland Brendel  
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.

\*\*\*\*\*

Unicorn: Continual learning with a universal, off-policy agent  
Daniel J. Mankowitz, Augustin Židek, André Barreto, Dan Horgan, Matteo Hessel, John Quan, Junhyuk Oh, Hado van Hasselt, David Silver, Tom Schaul  
Some real-world domains are best characterized as a single task, but for others this perspective is limiting. Instead, some tasks continually grow in complexity, in tandem with the agent's competence. In continual learning there are no explicit task boundaries or curricula. As learning agents have become more powerful, continual learning remains one of the frontiers that has resisted quick progress. To test continual learning capabilities we consider a challenging 3D domain with an implicit sequence of tasks and sparse rewards. We propose a novel agent architecture called Unicorn, which demonstrates strong continual learning and outperforms several baseline agents on the proposed domain. The agent achieves this by jointly representing and efficiently learning multiple policies for multiple goals, using a parallel off-policy learning setup.

\*\*\*\*\*

On the Turing Completeness of Modern Neural Network Architectures  
Jorge Pérez, Javier Marinković, Pablo Barceló  
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 Trans

former (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

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.

\*\*\*\*\*

Denoise while Aggregating: Collaborative Learning in Open-Domain Question Answering

Haozhe Ji, Yankai Lin, Zhiyuan Liu, Maosong Sun

The open-domain question answering (OpenQA) task aims to extract answers that match specific questions from a distantly supervised corpus. Unlike supervised reading comprehension (RC) datasets where questions are designed for particular paragraphs, background sentences in OpenQA datasets are more prone to noise. We observe that most existing OpenQA approaches are vulnerable to noise since they simply regard those sentences that contain the answer span as ground truths and ignore the plausible correlation between the sentences and the question. To address this deficiency, we introduce a unified and collaborative model that leverages alignment information from query-sentence pairs in a small-scale supervised RC dataset and aggregates relevant evidence from distantly supervised corpus to answer open-domain questions. We evaluate our model on several real-world OpenQA datasets, and experimental results show that our collaborative learning methods outperform the existing baselines significantly.

\*\*\*\*\*

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

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, Richard Socher

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.

\*\*\*\*\*

Three continual learning scenarios and a case for generative replay

Gido M. van de Ven, Andreas S. Tolias

Standard artificial neural networks suffer from the well-known issue of catastrophic forgetting, making continual or lifelong learning problematic. Recently, numerous methods have been proposed for continual learning, but due to differences in evaluation protocols it is difficult to directly compare their performance. To enable more meaningful comparisons, we identified three distinct continual learning scenarios based on whether task identity is known and, if it is not, whether it needs to be inferred. Performing the split and permuted MNIST task protocols according to each of these scenarios, we found that regularization-based approaches (e.g., elastic weight consolidation) failed when task identity needed to be inferred. In contrast, generative replay combined with distillation (i.e., using class probabilities as "soft targets") achieved superior performance in all three scenarios. In addition, we reduced the computational cost of generative replay by integrating the generative model into the main model.

\*\*\*\*\*

On the Relationship between Neural Machine Translation and Word Alignment

Xintong Li, Lemao Liu, Guanlin Li, Max Meng, Shuming Shi

Prior researches suggest that attentional neural machine translation (NMT) is able to capture word alignment by attention, however, to our surprise, it almost fails for NMT models with multiple attentional layers except for those with a single layer. This paper introduces two methods to induce word alignment from general neural machine translation models. Experiments verify that both methods obtain much better word alignment than the method by attention. Furthermore, based on one of the proposed methods, we design a criterion to divide target words into two categories (i.e. those mostly contributed from source "CFS" words and the other words mostly contributed from target "CFT" words), and analyze word alignment under these two categories in depth. We find that although NMT models are difficult to capture word alignment for CFT words but these words do not sacrifice translation quality significantly, which provides an explanation why NMT is more successful for translation yet worse for word alignment compared to statistical machine translation. We further demonstrate that word alignment errors for CFS words are responsible for translation errors in some extent by measuring the correlation between word alignment and translation for several NMT systems.

\*\*\*\*\*

Polar Prototype Networks

Pascal Mettes, Elise van der Pol, Cees G. M. Snoek

This paper proposes a neural network for classification and regression, without the need to learn layout structures in the output space. Standard solutions such as softmax cross-entropy and mean squared error are effective but parametric, meaning that known inductive structures such as maximum margin separation and simplicity (Occam's Razor) need to be learned for the task at hand. Instead, we propose polar prototype networks, a class of networks that explicitly states the structure, i.e. the layout, of the output. The structure is defined by polar prototypes, points on the hypersphere of the output space. For classification, each class is described by a single polar prototype and they are a priori distributed with maximal separation and equal shares on the hypersphere. Classes are assigned to prototypes randomly or based on semantic priors and training becomes a matter of minimizing angular distances between examples and their class prototypes. For regression, we show that training can be performed as a polar interpolation between two prototypes, arriving at a regression with higher-dimensional outputs. From empirical analysis, we find that polar prototype networks benefit from large margin separation and semantic class structure, while only requiring a minimal amount of output dimensions. While the structure is simple, the performance is on par with (classification) or better than (regression) standard network methods. Moreover, we show that we gain the ability to perform regression and classification jointly in the same space, which is disentangled and interpretable by design.

\*\*\*\*\*

DANA: Scalable Out-of-the-box Distributed ASGD Without Retuning

Ido Hakimi, Saar Barkai, Moshe Gabel, Assaf Schuster

Distributed computing can significantly reduce the training time of neural networks. Despite its potential, however, distributed training has not been widely adopted: scaling the training process is difficult, and existing SGD methods require substantial tuning of hyperparameters and learning schedules to achieve sufficient accuracy when increasing the number of workers. In practice, such tuning can be prohibitively expensive given the huge number of potential hyperparameter configurations and the effort required to test each one.

We propose DANA, a novel approach that scales out-of-the-box to large clusters using the same hyperparameters and learning schedule optimized for training on a single worker, while maintaining similar final accuracy without additional overhead. DANA estimates the future value of model parameters by adapting Nesterov Accelerated Gradient to a distributed setting, and so mitigates the effect of gradient staleness, one of the main difficulties in scaling SGD to more workers.

Evaluation on three state-of-the-art network architectures and three datasets shows that DANA scales as well as or better than existing work without having to tune any hyperparameters or tweak the learning schedule. For example, DANA achieves 75.73% accuracy on ImageNet when training ResNet-50 with 16 workers, similar to the non-distributed baseline.

\*\*\*\*\*

iRDA Method for Sparse Convolutional Neural Networks

Xiaodong Jia, Liang Zhao, Lian Zhang, Juncai He, Jinchao Xu

We propose a new approach, known as the iterative regularized dual averaging (iRDA), to improve the efficiency of convolutional neural networks (CNN) by significantly reducing the redundancy of the model without reducing its accuracy. The method has been tested for various data sets, and proven to be significantly more efficient than most existing compressing techniques in the deep learning literature. For many popular data sets such as MNIST and CIFAR-10, more than 95% of the weights can be zeroed out without losing accuracy. In particular, we are able to make ResNet18 with 95% sparsity to have an accuracy that is comparable to that of a much larger model ResNet50 with the best 60% sparsity as reported in the literature.

\*\*\*\*\*

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

Victor Zhong, Caiming Xiong, Nitish Shirish Keskar, Richard Socher

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.

\*\*\*\*\*

Rectified Gradient: Layer-wise Thresholding for Sharp and Coherent Attribution Maps

Beomsu Kim, Junghoon Seo, Jeongyeol Choe, Jamyoungh Koo, Seunghyeon Jeon, Taegyun Jeon  
Saliency map, or the gradient of the score function with respect to the input, is the most basic means of interpreting deep neural network decisions. However, saliency maps are often visually noisy. Although several hypotheses were proposed to account for this phenomenon, there is no work that provides a rigorous analysis



sis of noisy saliency maps. This may be a problem as numerous advanced attribution methods were proposed under the assumption that the existing hypotheses are true. In this paper, we identify the cause of noisy saliency maps. Then, we propose Rectified Gradient, a simple method that significantly improves saliency maps by alleviating that cause. Experiments showed effectiveness of our method and its superiority to other attribution methods. Codes and examples for the experiments will be released in public.

\*\*\*\*\*

#### Near-Optimal Representation Learning for Hierarchical Reinforcement Learning

Ofir Nachum, Shixiang Gu, Honglak Lee, Sergey Levine

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.

\*\*\*\*\*

#### Projective Subspace Networks For Few-Shot Learning

Christian Simon, Piotr Koniusz, Mehrtash Harandi

Generalization from limited examples, usually studied under the umbrella of meta-learning, equips learning techniques with the ability to adapt quickly in dynamical environments and proves to be an essential aspect of lifelong learning. In this paper, we introduce the Projective Subspace Networks (PSN), a deep learning paradigm that learns non-linear embeddings from limited supervision. In contrast to previous studies, the embedding in PSN deems samples of a given class to form an affine subspace. We will show that such modeling leads to robust solutions, yielding competitive results on supervised and semi-supervised few-shot classification. Moreover, our PSN approach has the ability of end-to-end learning. In contrast to previous works, our projective subspace can be thought of as a richer representation capturing higher-order information datapoints for modeling new concepts.

\*\*\*\*\*

#### Assumption Questioning: Latent Copying and Reward Exploitation in Question Generation

Tom Hosking, Sebastian Riedel

Question generation is an important task for improving our ability to process natural language data, with additional challenges over other sequence transformation tasks. Recent approaches use modifications to a Seq2Seq architecture inspired by advances in machine translation, but unlike translation the input and output vocabularies overlap significantly, and there are many different valid questions for each input. Approaches using copy mechanisms and reinforcement learning have shown promising results, but there are ambiguities in the exact implementation that have not yet been investigated. We show that by removing inductive bias from the model and allowing the choice of generation path to become latent, we achieve substantial improvements over implementations biased with both naive and smart heuristics. We perform a human evaluation to confirm these findings. We show that although policy gradient methods may be used to decouple training from the ground truth and optimise directly for quality metrics that have previously been assumed to be good choices, these objectives are poorly aligned with human judgement and the model simply learns to exploit the weaknesses of the reward source. Finally, we show that an adversarial objective learned directly from the ground truth data is not able to generate a useful training signal.

\*\*\*\*\*

#### Pixel Redrawn For A Robust Adversarial Defense

Jiacang Ho,Dae-Ki Kang

Recently, an adversarial example becomes a serious problem to be aware of because it can fool trained neural networks easily.

To prevent the issue, many researchers have proposed several defense techniques such as adversarial training, input transformation, stochastic activation pruning, etc.

In this paper, we propose a novel defense technique, Pixel Redrawn (PR) method, which redraws every pixel of training images to convert them into distorted images.

The motivation for our PR method is from the observation that the adversarial attacks have redrawn some pixels of the original image with the known parameters of the trained neural network.

Mimicking these attacks, our PR method redraws the image without any knowledge of the trained neural network.

This method can be similar to the adversarial training method but our PR method can be used to prevent future attacks.

Experimental results on several benchmark datasets indicate our PR method not only relieves the over-fitting issue when we train neural networks with a large number of epochs, but it also boosts the robustness of the neural network.

\*\*\*\*\*

Execution-Guided Neural Program Synthesis

Xinyun Chen,Chang Liu,Dawn Song

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 Speichner,Armin B. Cremers,Joerg Zimmermann

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}](https://github.com/gnodisnait/nball4tree.git) and [\url{https://github.com/gnodisnait/bp94nball.git}](https://github.com/gnodisnait/bp94nball.git).

\*\*\*\*\*

Adaptive Convolutional Neural Networks

Julio Cesar Zamora,Jesus Adan Cruz Vargas,Omesh Tickoo

The quest for increased visual recognition performance has led to the development of highly complex neural networks with very deep topologies. To avoid high com

puting resource requirements of such complex networks and to enable operation on devices with limited resources, this paper introduces adaptive kernels for convolutional layers. Motivated by the non-linear perception response in human visual cells, the input image is used to define the weights of a dynamic kernel called Adaptive kernel. This new adaptive kernel is used to perform a second convolution of the input image generating the output pixel. Adaptive kernels enable accurate recognition with lower memory requirements; This is accomplished through reducing the number of kernels and the number of layers needed in the typical CNN configuration, in addition to reducing the memory used, increasing 2X the training speed and the number of activation function evaluations. Our experiments show a reduction of 70X in the memory used for MNIST, maintaining 99% accuracy and 16X memory reduction for CIFAR10 with 92.5% accuracy.

\*\*\*\*\*

#### Accelerated Value Iteration via Anderson Mixing

Yujun Li, Chengzhuo Ni, Guangzeng Xie, Wenhao Yang, Shuchang Zhou, Zhihua Zhang

Acceleration for reinforcement learning methods is an important and challenging theme. We introduce the Anderson acceleration technique into the value iteration, developing an accelerated value iteration algorithm that we call Anderson Accelerated Value Iteration (A2VI). We further apply our method to the Deep Q-learning algorithm, resulting in the Deep Anderson Accelerated Q-learning (DA2Q) algorithm. Our approach can be viewed as an approximation of the policy evaluation by interpolating on historical data. A2VI is more efficient than the modified policy iteration, which is a classical approximate method for policy evaluation. We give a theoretical analysis of our algorithm and conduct experiments on both toy problems and Atari games. Both the theoretical and empirical results show the effectiveness of our algorithm.

\*\*\*\*\*

#### SHAMANN: Shared Memory Augmented Neural Networks

Cosmin I. Bercea, Olivier Pauly, Andreas K. Maier, Florin C. Ghesu

Current state-of-the-art methods for semantic segmentation use deep neural networks to learn the segmentation mask from the input image signal as an image-to-image mapping. While these methods effectively exploit global image context, the learning and computational complexities are high. We propose shared memory augmented neural network actors as a dynamically scalable alternative. Based on a decomposition of the image into a sequence of local patches, we train such actors to sequentially segment each patch. To further increase the robustness and better capture shape priors, an external memory module is shared between different actors, providing an implicit mechanism for image information exchange. Finally, the patch-wise predictions are aggregated to a complete segmentation mask. We demonstrate the benefits of the new paradigm on a challenging lung segmentation problem based on chest X-Ray images, as well as on two synthetic tasks based on the MNIST dataset. On the X-Ray data, our method achieves state-of-the-art accuracy with a significantly reduced model size compared to reference methods. In addition, we reduce the number of failure cases by at least half.

\*\*\*\*\*

#### FROM DEEP LEARNING TO DEEP DEDUCING: AUTOMATICALLY TRACKING DOWN NASH EQUILIBRIUM THROUGH AUTONOMOUS NEURAL AGENT, A POSSIBLE MISSING STEP TOWARD GENERAL A.I.

Brown Wang

Contrary to most reinforcement learning studies, which emphasize on training a deep neural network to approximate its output layer to certain strategies, this paper proposes a reversed method for reinforcement learning. We call this "Deep Deducing". In short, after adequately training a deep neural network according to a strategy-environment-to-payoff table, then we initialize randomized strategy input and propagate the error between the actual output and the desired output back to the initially-randomized strategy input in the "input layer" of the trained deep neural network gradually to perform a task similar to "human deduction". And we view the final strategy input in the "input layer" as the fittest strategy for a neural network when confronting the observed environment input from the world outside.

\*\*\*\*\*

Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics

Antonin Raffin, Ashley Hill, René Traoré, Timothée Lesort, Natalia Díaz-Rodríguez, David Filliat

Scaling end-to-end reinforcement learning to control real robots from vision presents a series of challenges, in particular in terms of sample efficiency. Against end-to-end learning, state representation learning can help learn a compact, efficient and relevant representation of states that speeds up policy learning, reducing the number of samples needed, and that is easier to interpret. We evaluate several state representation learning methods on goal based robotics tasks and propose a new unsupervised model that stacks representations and combines strengths of several of these approaches. This method encodes all the relevant features, performs on par or better than end-to-end learning, and is robust to hyper-parameters change.

\*\*\*\*\*

Integrated Steganography and Steganalysis with Generative Adversarial Networks

Chong Yu

Recently, generative adversarial network is the hotspot in research areas and industrial application areas. It's application on data generation in computer vision is most common usage. This paper extends its application to data hiding and security area. In this paper, we propose the novel framework to integrate steganography and steganalysis processes. The proposed framework applies generative adversarial networks as the core structure. The discriminative model simulate the steganalysis process, which can help us understand the sensitivity of cover images to semantic changes. The steganography generative model is to generate stego image which is aligned with the original cover image, and attempts to confuse steganalysis discriminative model. The introduction of cycle discriminative model and inconsistent loss can help to enhance the quality and security of generated stego image in the iterative training process. Training dataset is mixed with intact images as well as intentional attacked images. The mix training process can further improve the robustness and security of new framework. Through the qualitative, quantitative experiments and analysis, this novel framework shows compelling performance and advantages over the current state-of-the-art methods in steganography and steganalysis benchmarks.

\*\*\*\*\*

Learning Kolmogorov Models for Binary Random Variables

Hadi Ghauch, Hossein S. Ghadikolaei, Mikael Skoglund, Carlo Fischione

We propose a framework for learning a Kolmogorov model, for a collection of binary random variables. More specifically, we derive conditions that link (in the sense of implications in mathematical logic) outcomes of specific random variables and extract valuable relations from the data. We also propose an efficient algorithm for computing the model and show its first-order optimality, despite the combinatorial nature of the learning problem. We exemplify our general framework to recommendation systems and gene expression data. We believe that the work is a significant step toward interpretable machine learning.

\*\*\*\*\*

DelibGAN: Coarse-to-Fine Text Generation via Adversarial Network

Ke Wang, Xiaojun Wan

In this paper, we propose a novel adversarial learning framework, namely DelibGAN, for generating high-quality sentences without supervision. Our framework consists of a coarse-to-fine generator, which contains a first-pass decoder and a second-pass decoder, and a multiple instance discriminator. And we propose two training mechanisms DelibGAN-I and DelibGAN-II. The discriminator is used to fine-tune the second-pass decoder in DelibGAN-I and further evaluate the importance of each word and tune the first-pass decoder in DelibGAN-II. We compare our models with several typical and state-of-the-art unsupervised generic text generation models on three datasets (a synthetic dataset, a descriptive text dataset and a sentimental text dataset). Both qualitative and quantitative experimental results show that our models produce more realistic samples, and DelibGAN-II performs best.

\*\*\*\*\*

## Consistency-based anomaly detection with adaptive multiple-hypotheses predictions

Duc Tam Nguyen, Zhongyu Lou, Michael Klar, Thomas Brox

In one-class-learning tasks, only the normal case can be modeled with data, whereas the variation of all possible anomalies is too large to be described sufficiently by samples. Thus, due to the lack of representative data, the wide-spread discriminative approaches cannot cover such learning tasks, and rather generative models, which attempt to learn the input density of the normal cases, are used. However, generative models suffer from a large input dimensionality (as in images) and are typically inefficient learners. We propose to learn the data distribution more efficiently with a multi-hypotheses autoencoder. Moreover, the model is criticized by a discriminator, which prevents artificial data modes not supported by data, and which enforces diversity across hypotheses. This consistency-based anomaly detection (ConAD) framework allows the reliable identification of out-of-distribution samples. For anomaly detection on CIFAR-10, it yields up to 3.9% points improvement over previously reported results. On a real anomaly detection task, the approach reduces the error of the baseline models from 6.8% to 1.5%.

\*\*\*\*\*

## A quantifiable testing of global translational invariance in Convolutional and Capsule Networks

Weikai Qi

We design simple and quantifiable testing of global translation-invariance in deep learning models trained on the MNIST dataset. Experiments on convolutional and capsule neural networks show that both models have poor performance in dealing with global translation-invariance; however, the performance improved by using data augmentation. Although the capsule network is better on the MNIST testing dataset, the convolutional neural network generally has better performance on the translation-invariance.

\*\*\*\*\*

## RedSync : Reducing Synchronization Traffic for Distributed Deep Learning

Jiarui Fang, Cho-Jui Hsieh

Data parallelism has become a dominant method to scale Deep Neural Network (DNN) training across multiple nodes. Since the synchronization of the local models or gradients can be a bottleneck for large-scale distributed training, compressing communication traffic has gained widespread attention recently. Among several recent proposed compression algorithms, Residual Gradient Compression (RGC) is one of the most successful approaches---it can significantly compress the transmitting message size (0.1% of the gradient size) of each node and still preserve accuracy. However, the literature on compressing deep networks focuses almost exclusively on achieving good compression rate, while the efficiency of RGC in real implementation has been less investigated. In this paper, we develop an RGC method that achieves significant training time improvement in real-world multi-GPU systems. Our proposed RGC system design called RedSync, introduces a set of optimizations to reduce communication bandwidth while introducing limited overhead. We examine the performance of RedSync on two different multiple GPU platforms, including a supercomputer and a multi-card server. Our test cases include image classification on Cifar10 and ImageNet, and language modeling tasks on Penn Treebank and Wiki2 datasets. For DNNs featured with high communication to computation ratio, which has long been considered with poor scalability, RedSync shows significant performance improvement.

\*\*\*\*\*

## EMI: Exploration with Mutual Information Maximizing State and Action Embeddings

Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, Hyun Oh Song

Policy optimization struggles when the reward feedback signal is very sparse and essentially becomes a random search algorithm until the agent stumbles upon a rewarding or the goal state. Recent works utilize intrinsic motivation to guide the exploration via generative models, predictive forward models, or more ad-hoc measures of surprise. We propose EMI, which is an exploration method that constr

ucts embedding representation of states and actions that does not rely on generative decoding of the full observation but extracts predictive signals that can be used to guide exploration based on forward prediction in the representation space. Our experiments show the state of the art performance on challenging locomotion task with continuous control and on image-based exploration tasks with discrete actions on Atari.

\*\*\*\*\*

A Multi-modal one-class generative adversarial network for anomaly detection in manufacturing

Shuhui Qu, Janghwan Lee, Wei Xiong, Wonhyouk Jang, Jie Wang

One class anomaly detection on high-dimensional data is one of the critical issues in both fundamental machine learning research area and manufacturing applications. A good anomaly detection should accurately discriminate anomalies from normal data. Although most previous anomaly detection methods achieve good performances, they do not perform well on high-dimensional imbalanced data- set 1) with a limited amount of data; 2) multi-modal distribution; 3) few anomaly data. In this paper, we develop a multi-modal one-class generative adversarial network based detector (MMOC-GAN) to distinguish anomalies from normal data (products). Apart from a domain-specific feature extractor, our model leverage a generative adversarial network (GAN). The generator takes in a modified noise vector using a pseudo latent prior and generate samples at the low-density area of the given normal data to simulate the anomalies. The discriminator then is trained to distinguish the generated samples from the normal samples. Since the generated samples simulate the low density area for each modal, the discriminator could directly detect anomalies from normal data. Experiments demonstrate that our model outperforms the state-of-the-art one-class classification models and other anomaly detection methods on both normal data and anomalies accuracy, as well as the F1 score. Also, the generated samples can fully capture the low density area of different types of products.

\*\*\*\*\*

Pumpout: A Meta Approach for Robustly Training Deep Neural Networks with Noisy Labels

Bo Han, Gang Niu, Jiangchao Yao, Xingrui Yu, Miao Xu, Ivor Tsang, Masashi Sugiyama

It is challenging to train deep neural networks robustly on the industrial-level data, since labels of such data are heavily noisy, and their label generation processes are normally agnostic. To handle these issues, by using the memorization effects of deep neural networks, we may train deep neural networks on the whole dataset only the first few iterations. Then, we may employ early stopping or the small-loss trick to train them on selected instances. However, in such training procedures, deep neural networks inevitably memorize some noisy labels, which will degrade their generalization. In this paper, we propose a meta algorithm called Pumpout to overcome the problem of memorizing noisy labels. By using scaled stochastic gradient ascent, Pumpout actively squeezes out the negative effects of noisy labels from the training model, instead of passively forgetting these effects. We leverage Pumpout to upgrade two representative methods: MentorNet and Backward Correction. Empirical results on benchmark vision and text datasets demonstrate that Pumpout can significantly improve the robustness of representative methods.

\*\*\*\*\*

Incremental training of multi-generative adversarial networks

Qi Tan, Pingzhong Tang, Ke Xu, Weiran Shen, Song Zuo

Generative neural networks map a standard, possibly distribution to a complex high-dimensional distribution, which represents the real world data set. However, a determinate input distribution as well as a specific architecture of neural networks may impose limitations on capturing the diversity in the high dimensional target space. To resolve this difficulty, we propose a training framework that greedily produce a series of generative adversarial networks that incrementally capture the diversity of the target space. We show theoretically and empirically that our training algorithm converges to the theoretically optimal distribution

, the projection of the real distribution onto the convex hull of the network's distribution space.

\*\*\*\*\*

S3TA: A Soft, Spatial, Sequential, Top-Down Attention Model

Alex Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, Danilo J. Rezende

We present a soft, spatial, sequential, top-down attention model (S3TA). This model uses a soft attention mechanism to bottleneck its view of the input. A recurrent core is used to generate query vectors, which actively select information from the input by correlating the query with input- and space-dependent key maps at different spatial locations.

We demonstrate the power and interpretability of this model under two settings. First, we build an agent which uses this attention model in RL environments and show that we can achieve performance competitive with state-of-the-art models while producing attention maps that elucidate some of the strategies used to solve the task. Second, we use this model in supervised learning tasks and show that it also achieves competitive performance and provides interpretable attention maps that show some of the underlying logic in the model's decision making.

\*\*\*\*\*

ROBUST ESTIMATION VIA GENERATIVE ADVERSARIAL NETWORKS

Chao GAO, jiyi LIU, Yuan YAO, Weizhi ZHU

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 network structure in the discriminator class are shown to be necessary for robust estimation.

\*\*\*\*\*

Learning Discriminators as Energy Networks in Adversarial Learning

Pingbo Pan, Yan Yan, Tianbao Yang, Yi Yang

We propose a novel adversarial learning framework in this work. Existing adversarial learning methods involve two separate networks, i.e., the structured prediction models and the discriminative models, in the training. The information captured by discriminative models complements that in the structured prediction models, but few existing researches have studied on utilizing such information to improve structured prediction models at the inference stage. In this work, we propose to refine the predictions of structured prediction models by effectively integrating discriminative models into the prediction. Discriminative models are treated as energy-based models. Similar to the adversarial learning, discriminative models are trained to estimate scores which measure the quality of predicted outputs, while structured prediction models are trained to predict contrastive outputs with maximal energy scores. In this way, the gradient vanishing problem is ameliorated, and thus we are able to perform inference by following the ascent gradient directions of discriminative models to refine structured prediction models. The proposed method is able to handle a range of tasks, \emph{e.g.}, multi-label classification and image segmentation. Empirical results on these two tasks validate the effectiveness of our learning method.

\*\*\*\*\*

Dynamic Pricing on E-commerce Platform with Deep Reinforcement Learning

Jiaxi Liu, Yidong Zhang, Xiaoqing Wang, Yuming Deng, Xingyu Wu, Miaolan Xie

In this paper we develop an approach based on deep reinforcement learning (DRL) to address dynamic pricing problem on E-commerce platform. We models real-world

E-commerce dynamic pricing problem as Markov Decision Process. Environment state are defined with four groups of different business data. We make several main improvements on the state-of-the-art DRL-based dynamic pricing approaches: 1. We first extend the application of dynamic pricing to a continuous pricing action space. 2. We solve the unknown demand function problem by designing different reward functions. 3. The cold-start problem is addressed by introducing pre-training and evaluation using the historical sales data. Field experiments are designed and conducted on real-world E-commerce platform, pricing thousands of SKUs of products lasting for months. The experiment results shows that, on E-commerce platform, the difference of the revenue conversion rates (DRCR) is a more suitable reward function than the revenue only, which is different from the conclusion from previous researches. Meanwhile, the proposed continuous action model performs better than the discrete one.

\*\*\*\*\*

Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search

Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau

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 as Stochastic Value Gradient can be interpreted as counterfactual methods.

\*\*\*\*\*

LEARNING ADVERSARIAL EXAMPLES WITH RIEMANNIAN GEOMETRY

Shufei Zhang, Kaizhu Huang, Rui Zhang, Amir Hussain

Adversarial examples, referred to as augmented data points generated by imperceptible perturbation of input samples, have recently drawn much attention. Well-crafted adversarial examples may even mislead state-of-the-art deep models to make wrong predictions easily. To alleviate this problem, many studies focus on investigating how adversarial examples can be generated and/or resisted. All the existing work handles this problem in the Euclidean space, which may however be unable to describe data geometry. In this paper, we propose a generalized framework that addresses the learning problem of adversarial examples with Riemannian geometry. Specifically, we define the local coordinate systems on Riemannian manifold, develop a novel model called Adversarial Training with Riemannian Manifold, and design a series of theory that manages to learn the adversarial examples in the Riemannian space feasibly and efficiently. The proposed work is important in that (1) it is a generalized learning methodology since Riemannian manifold space would be degraded to the Euclidean space in a special case; (2) it is the first work to tackle the adversarial example problem tractably through the perspective of geometry; (3) from the perspective of geometry, our method leads to the steepest direction of the loss function. We also provide a series of theory showing that our proposed method can truly find the decent direction for the loss function with a comparable computational time against traditional adversarial methods. Finally, the proposed framework demonstrates superior performance to the tra



ditional counterpart methods on benchmark data including MNIST, CIFAR-10 and SVHN.

\*\*\*\*\*

#### Robust Conditional Generative Adversarial Networks

Grigorios G. Chrysos, Jean Kossaifi, Stefanos Zafeiriou

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 shares 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

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.

\*\*\*\*\*

#### Pyramid Recurrent Neural Networks for Multi-Scale Change-Point Detection

Zahra Ebrahimzadeh, Min Zheng, Selcuk Karakas, Samantha Kleinberg

Many real-world time series, such as in activity recognition, finance, or climate science, have changepoints where the system's structure or parameters change. Detecting changes is important as they may indicate critical events. However, existing methods for changepoint detection face challenges when (1) the patterns of change cannot be modeled using simple and predefined metrics, and (2) changes can occur gradually, at multiple time-scales. To address this, we show how changepoint detection can be treated as a supervised learning problem, and propose a new deep neural network architecture that can efficiently identify both abrupt and gradual changes at multiple scales. Our proposed method, pyramid recurrent neural network (PRNN), is designed to be scale-invariant, by incorporating wavelets and pyramid analysis techniques from multi-scale signal processing. Through experiments on synthetic and real-world datasets, we show that PRNN can detect abrupt and gradual changes with higher accuracy than the state of the art and can extrapolate to detect changepoints at novel timescales that have not been seen in training.

\*\*\*\*\*

#### Visual Reasoning by Progressive Module Networks

Seung Wook Kim, Makarand Tapaswi, Sanja Fidler

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 ea

ch 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 performance 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.

\*\*\*\*\*

#### Large-Scale Visual Speech Recognition

Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Ben Coppin, Ben Laurie, Andrew Senior, Nando de Freitas

This work presents a scalable solution to continuous visual speech recognition. To achieve this, we constructed the largest existing visual speech recognition dataset, consisting of pairs of text and video clips of faces speaking (3,886 hours of video). In tandem, we designed and trained an integrated lipreading system, consisting of a video processing pipeline that maps raw video to stable videos of lips and sequences of phonemes, a scalable deep neural network that maps the lip videos to sequences of phoneme distributions, and a production-level speech decoder that outputs sequences of words. The proposed system achieves a word error rate (WER) of 40.9% as measured on a held-out set. In comparison, professional lipreaders achieve either 86.4% or 92.9% WER on the same dataset when having access to additional types of contextual information. Our approach significantly improves on previous lipreading approaches, including variants of LipNet and of Watch, Attend, and Spell (WAS), which are only capable of 89.8% and 76.8% WER respectively.

\*\*\*\*\*

#### Real-time Neural-based Input Method

Jiali Yao, Raphael Shu, Xinjian Li, Katsutoshi Ohtsuki, Hideki Nakayama

The input method is an essential service on every mobile and desktop devices that provides text suggestions. It converts sequential keyboard inputs to the characters in its target language, which is indispensable for Japanese and Chinese users. Due to critical resource constraints and limited network bandwidth of the target devices, applying neural models to input method is not well explored. In this work, we apply a LSTM-based language model to input method and evaluate its performance for both prediction and conversion tasks with Japanese BCCWJ corpus.

We articulate the bottleneck to be the slow softmax computation during conversion. To solve the issue, we propose incremental softmax approximation approach, which computes softmax with a selected subset vocabulary and fix the stale probabilities when the vocabulary is updated in future steps. We refer to this method as incremental selective softmax. The results show a two order speedup for the softmax computation when converting Japanese input sequences with a large vocabulary, reaching real-time speed on commodity CPU. We also exploit the model compressing potential to achieve a 92% model size reduction without losing accuracy.

\*\*\*\*\*

#### Hindsight policy gradients

Paulo Rauber, Avinash Ummadisingu, Filipe Mutz, Jürgen Schmidhuber

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.

\*\*\*\*\*

#### Multi-Scale Stacked Hourglass Network for Human Pose Estimation

Chunsheng Guo, Wenlong Du, Na Ying

Stacked hourglass network has become an important model for Human pose estimation. The estimation of human body posture depends on the global information of the keypoints type and the local information of the keypoints location. The consistent processing of inputs and constraints makes it difficult to form differentiated and determined collaboration mechanisms for each stacked hourglass network. In this paper, we propose a Multi-Scale Stacked Hourglass (MSSH) network to highlight the differentiation capabilities of each Hourglass network for human pose estimation. The pre-processing network forms feature maps of different scales, and dispatch them to various locations of the stack hourglass network, where the small-scale features reach the front of stacked hourglass network, and large-scale features reach the rear of stacked hourglass network. And a new loss function is proposed for multi-scale stacked hourglass network. Different keypoints have different weight coefficients of loss function at different scales, and the keypoints weight coefficients are dynamically adjusted from the top-level hourglass network to the bottom-level hourglass network. Experimental results show that the proposed method is competitive with respect to the comparison algorithm on MPII and LSP datasets.

\*\*\*\*\*

#### Computation-Efficient Quantization Method for Deep Neural Networks

Parichay Kapoor, Dongsoo Lee, Byeongwook Kim, Saehyung Lee

Deep Neural Networks, being memory and computation intensive, are a challenge to deploy in smaller devices. Numerous quantization techniques have been proposed to reduce the inference latency/memory consumption. However, these techniques impose a large overhead on the training procedure or need to change the training process. We present a non-intrusive quantization technique based on re-training the full precision model, followed by directly optimizing the corresponding binary model. The quantization training process takes no longer than the original training process. We also propose a new loss function to regularize the weights, resulting in reduced quantization error. Combining both help us achieve full precision accuracy on CIFAR dataset using binary quantization. We also achieve full precision accuracy on WikiText-2 using 2 bit quantization. Comparable results are also shown for ImageNet. We also present a 1.5 bits hybrid model exceeding the performance of TWN LSTM model for WikiText-2.

\*\*\*\*\*

#### Modular Deep Probabilistic Programming

Zhenwen Dai, Eric Meissner, Neil D. Lawrence

Modularity is a key feature of deep learning libraries but has not been fully exploited for probabilistic programming. We propose to improve modularity of probabilistic programming language by offering not only plain probabilistic distributions but also sophisticated probabilistic model such as Bayesian non-parametric models as fundamental building blocks. We demonstrate this idea by presenting a modular probabilistic programming language MXFusion, which includes a new type of re-usable building blocks, called probabilistic modules. A probabilistic module consists of a set of random variables with associated probabilistic distributions and dedicated inference methods. Under the framework of variational inference, the pre-specified inference methods of individual probabilistic modules can be transparently used for inference of the whole probabilistic model. We demonstrate the power and convenience of probabilistic modules in MXFusion with various examples of Gaussian process models, which are evaluated with experiments on real data.

\*\*\*\*\*

#### Efficient Convolutional Neural Network Training with Direct Feedback Alignment

Donghyeon Han, Hoi-jun Yoo

There were many algorithms to substitute the back-propagation (BP) in the deep n

neural network (DNN) training. However, they could not become popular because their training accuracy and the computational efficiency were worse than BP. One of them was direct feedback alignment (DFA), but it showed low training performance especially for the convolutional neural network (CNN). In this paper, we overcome the limitation of the DFA algorithm by combining with the conventional BP during the CNN training. To improve the training stability, we also suggest the feedback weight initialization method by analyzing the patterns of the fixed random matrices in the DFA. Finally, we propose the new training algorithm, binary direct feedback alignment (BDFA) to minimize the computational cost while maintaining the training accuracy compared with the DFA. In our experiments, we use the CIFAR-10 and CIFAR-100 dataset to simulate the CNN learning from the scratch and apply the BDFA to the online learning based object tracking application to examine the training in the small dataset environment. Our proposed algorithms show better performance than conventional BP in both two different training tasks especially when the dataset is small.

\*\*\*\*\*

Better Accuracy with Quantified Privacy: Representations Learned via Reconstructive Adversarial Network

Sicong Liu, Anshumali Shrivastava, Junzhao Du, Lin Zhong

The remarkable success of machine learning, especially deep learning, has produced a variety of cloud-based services for mobile users. Such services require an end user to send data to the service provider, which presents a serious challenge to end-user privacy. To address this concern, prior works either add noise to the data or send features extracted from the raw data. They struggle to balance between the utility and privacy because added noise reduces utility and raw data can be reconstructed from extracted features.

This work represents a methodical departure from prior works: we balance between a measure of privacy and another of utility by leveraging adversarial learning to find a sweeter tradeoff. We design an encoder that optimizes against the reconstruction error (a measure of privacy), adversarially by a Decoder, and the inference accuracy (a measure of utility) by a Classifier. The result is RAN, a novel deep model with a new training algorithm that automatically extracts features for classification that are both private and useful.

It turns out that adversarially forcing the extracted features to only convey the intended information required by classification leads to an implicit regularization leading to better classification accuracy than the original model which completely ignores privacy. Thus, we achieve better privacy with better utility, a surprising possibility in machine learning! We conducted extensive experiments on five popular datasets over four training schemes, and demonstrate the superiority of RAN compared with existing alternatives.

\*\*\*\*\*

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

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

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.

\*\*\*\*\*

## Dirichlet Variational Autoencoder

Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon

This paper proposes Dirichlet Variational Autoencoder (DirVAE) using a Dirichlet prior for a continuous latent variable that exhibits the characteristic of the categorical probabilities. To infer the parameters of DirVAE, we utilize the stochastic gradient method by approximating the Gamma distribution, which is a component of the Dirichlet distribution, with the inverse Gamma CDF approximation. Additionally, we reshape the component collapsing issue by investigating two problem sources, which are decoder weight collapsing and latent value collapsing, and we show that DirVAE has no component collapsing; while Gaussian VAE exhibits the decoder weight collapsing and Stick-Breaking VAE shows the latent value collapsing. The experimental results show that 1) DirVAE models the latent representation result with the best log-likelihood compared to the baselines; and 2) DirVAE produces more interpretable latent values with no collapsing issues which the baseline models suffer from. Also, we show that the learned latent representation from the DirVAE achieves the best classification accuracy in the semi-supervised and the supervised classification tasks on MNIST, OMNIGLOT, and SVHN compared to the baseline VAEs. Finally, we demonstrated that the DirVAE augmented topic models show better performances in most cases.

\*\*\*\*\*

## Local Stability and Performance of Simple Gradient Penalty $\mu$ -Wasserstein GAN

Cheolhyeong Kim, Seungtae Park, Hyung Ju Hwang

Wasserstein GAN (WGAN) is a model that minimizes the Wasserstein distance between a data distribution and sample distribution. Recent studies have proposed stabilizing the training process for the WGAN and implementing the Lipschitz constraint. In this study, we prove the local stability of optimizing the simple gradient penalty  $\mu$ -WGAN (SGP  $\mu$ -WGAN) under suitable assumptions regarding the equilibrium and penalty measure  $\mu$ . The measure valued differentiation concept is employed to deal with the derivative of the penalty terms, which is helpful for handling abstract singular measures with lower dimensional support. Based on this analysis, we claim that penalizing the data manifold or sample manifold is the key to regularizing the original WGAN with a gradient penalty. Experimental results obtained with unintuitive penalty measures that satisfy our assumptions are also provided to support our theoretical results.

\*\*\*\*\*

## Why Do Neural Response Generation Models Prefer Universal Replies?

Bowen Wu, Nan Jiang, Zhifeng Gao, Zongsheng Wang, Suke Li, Wenge Rong, Baoxun Wang

Recent advances in neural Sequence-to-Sequence (Seq2Seq) models reveal a purely data-driven approach to the response generation task. Despite its diverse variants and applications, the existing Seq2Seq models are prone to producing short and generic replies, which blocks such neural network architectures from being uti

lized in practical open-domain response generation tasks. In this research, we analyze this critical issue from the perspective of the optimization goal of models and the specific characteristics of human-to-human conversational corpora. Our analysis is conducted by decomposing the goal of Neural Response Generation (NRG) into the optimizations of word selection and ordering. It can be derived from the decomposing that Seq2Seq based NRG models naturally tend to select common words to compose responses, and ignore the semantic of queries in word ordering.

On the basis of the analysis, we propose a max-marginal ranking regularization term to avoid Seq2Seq models from producing the generic and uninformative responses. The empirical experiments on benchmarks with several metrics have validated our analysis and proposed methodology.

\*\*\*\*\*

#### Optimal margin Distribution Network

Shen-Huan Lv, Lu Wang, Zhi-Hua Zhou

Recent research about margin theory has proved that maximizing the minimum margin like support vector machines does not necessarily lead to better performance, and instead, it is crucial to optimize the margin distribution. In the meantime, margin theory has been used to explain the empirical success of deep network in recent studies. In this paper, we present ODN (the Optimal margin Distribution Network), a network which embeds a loss function in regard to the optimal margin distribution. We give a theoretical analysis for our method using the PAC-Bayesian framework, which confirms the significance of the margin distribution for classification within the framework of deep networks. In addition, empirical results show that the ODN model always outperforms the baseline cross-entropy loss model consistently across different regularization situations. And our ODN model also outperforms the cross-entropy loss (Xent), hinge loss and soft hinge loss model in generalization task through limited training data.

\*\*\*\*\*

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

Peng Cao\*, Yilun Xu\*, Yuqing Kong, Yizhou Wang

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>.

\*\*\*\*\*

#### A Priori Estimates of the Generalization Error for Two-layer Neural Networks

Lei Wu, Chao Ma, Weinan E

New estimates for the generalization error are established for a nonlinear regression problem using a two-layer neural network model. These new estimates are a priori in nature in the sense that the bounds depend only on some norms of the underlying functions to be fitted, not the parameters in the model. In contrast, most existing results for neural networks are a posteriori in nature in the sense that the bounds depend on some norms of the model parameters. The error rates are comparable to that of the Monte Carlo method in terms of the size of the dataset. Moreover, these bounds are equally effective in the over-parametrized regime when the network size is much larger than the size of the dataset.

\*\*\*\*\*

## Characterizing the Accuracy/Complexity Landscape of Explanations of Deep Networks through Knowledge Extraction

Simon Odense, Artur d'Avila Garcez

Knowledge extraction techniques are used to convert neural networks into symbolic descriptions with the objective of producing more comprehensible learning models. The central challenge is to find an explanation which is more comprehensible than the original model while still representing that model faithfully. The distributed nature of deep networks has led many to believe that the hidden features of a neural network cannot be explained by logical descriptions simple enough to be understood by humans, and that compositional knowledge extraction should be abandoned in favour of other methods. In this paper we examine this question systematically by proposing a knowledge extraction method using  $M$ -of- $N$  rules which allows us to map the complexity/accuracy landscape of rules describing hidden features in a Convolutional Neural Network (CNN). Experiments reported in this paper show that the shape of this landscape reveals an optimal trade off between comprehensibility and accuracy, showing that each latent variable has an optimal  $M$ -of- $N$  rule to describe its behaviour. We find that the rules with optimal tradeoff in the first and final layer have a high degree of explainability whereas the rules with the optimal tradeoff in the second and third layer are less explainable. The results shed light on the feasibility of rule extraction from deep networks, and point to the value of compositional knowledge extraction as a method of explainability.

\*\*\*\*\*

## VARIATIONAL SGD: DROPOUT, GENERALIZATION AND CRITICAL POINT AT THE END OF CONVEXITY

Michael Tetelman

The goal of the paper is to propose an algorithm for learning the most generalizable solution from given training data. It is shown that Bayesian approach leads to a solution that dependent on statistics of training data and not on particular

samples. The solution is stable under perturbations of training data because it is defined by an integral contribution of multiple maxima of the likelihood and not by a single global maximum. Specifically, the Bayesian probability distribution

of parameters (weights) of a probabilistic model given by a neural network is estimated via recurrent variational approximations. Derived recurrent update rules correspond to SGD-type rules for finding a minimum of an effective loss that is an average of an original negative log-likelihood over the Gaussian distributions of weights, which makes it a function of means and variances. The effective loss is convex for large variances and non-convex in the limit of small variances. Among stationary solutions of the update rules there are trivial solutions with zero variances at local minima of the original loss and a single non-trivial solution with finite variances that is a critical point at the end of convexity of the effective loss

in the mean-variance space. At the critical point both first- and second-order gradients of the effective loss w.r.t. means are zero. The empirical study confirms that the critical point represents the most generalizable solution. While the location of

the critical point in the weight space depends on specifics of the used probabilistic model some properties at the critical point are universal and model independent.

\*\*\*\*\*

## Hierarchical Visuomotor Control of Humanoids

Josh Merel, Arun Ahuja, Vu Pham, Saran Tunyasuvunakool, Siqi Liu, Dhruva Tirumala, Nicolas Heess, Greg Wayne

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

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.

\*\*\*\*\*

Why do deep convolutional networks generalize so poorly to small image transformations?

Aharon Azulay, Yair Weiss

Deep convolutional network architectures are often assumed to guarantee generalization for small image translations and deformations. In this paper we show that modern CNNs (VGG16, ResNet50, and InceptionResNetV2) can drastically change their output when an image is translated in the image plane by a few pixels, and that this failure of generalization also happens with other realistic small image transformations. Furthermore, we see these failures to generalize more frequently in more modern networks. We show that these failures are related to the fact that the architecture of modern CNNs ignores the classical sampling theorem so that generalization is not guaranteed. We also show that biases in the statistics of commonly used image datasets makes it unlikely that CNNs will learn to be invariant to these transformations. Taken together our results suggest that the performance of CNNs in object recognition falls far short of the generalization capabilities of humans.

\*\*\*\*\*

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

Alexander Shekhovtsov, Boris Flach

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.

\*\*\*\*\*



A preconditioned accelerated stochastic gradient descent algorithm

Alexandru Onose, Seyed Iman Mossavat, Henk-Jan H. Smilde

We propose a preconditioned accelerated stochastic gradient method suitable for large scale optimization. We derive sufficient convergence conditions for the minimization of convex functions using a generic class of diagonal preconditioners and provide a formal convergence proof based on a framework originally used for on-line learning. Inspired by recent popular adaptive per-feature algorithms, we propose a specific preconditioner based on the second moment of the gradient. The sufficient convergence conditions motivate a critical adaptation of the per-feature updates in order to ensure convergence. We show empirical results for the minimization of convex and non-convex cost functions, in the context of neural network training. The method compares favorably with respect to current, first order, stochastic optimization methods.

\*\*\*\*\*

Hierarchical Reinforcement Learning via Advantage-Weighted Information Maximization

Takayuki Osa, Voot Tangkaratt, Masashi Sugiyama

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.

\*\*\*\*\*

Formal Limitations on the Measurement of Mutual Information

David McAllester, Karl Stratos

Motivated by applications to unsupervised learning, we consider the problem of measuring mutual information. Recent analysis has shown that naive kNN estimators of mutual information have serious statistical limitations motivating more refined methods. In this paper we prove that serious statistical limitations are inherent to any measurement method. More specifically, we show that any distribution-free high-confidence lower bound on mutual information cannot be larger than  $O(\frac{1}{\ln N})$  where  $N$  is the size of the data sample. We also analyze the Donsker-Varadhan lower bound on KL divergence in particular and show that, when simple statistical considerations are taken into account, this bound can never produce a high-confidence value larger than  $\frac{1}{\ln N}$ . While large high-confidence lower bounds are impossible, in practice one can use estimators without formal guarantees. We suggest expressing mutual information as a difference of entropies and using cross entropy as an entropy estimator. We observe that, although cross entropy is only an upper bound on entropy, cross-entropy estimates converge to the true cross entropy at the rate of  $\frac{1}{\sqrt{N}}$ .

\*\*\*\*\*

Causal importance of orientation selectivity for generalization in image recognition

Jumpei Ukita

Although both our brain and deep neural networks (DNNs) can perform high-level sensory-perception tasks such as image or speech recognition, the inner mechanism of these hierarchical information-processing systems is poorly understood in both neuroscience and machine learning. Recently, Morcos et al. (2018) examined the effect of class-selective units in DNNs, i.e., units with high-level selectivity

ty, on network generalization, concluding that hidden units that are selectively activated by specific input patterns may harm the network's performance. In this study, we revisit their hypothesis, considering units with selectivity for lower-level features, and argue that selective units are not always harmful to the network performance. Specifically, by using DNNs trained for image classification (7-layer CNNs and VGG16 trained on CIFAR-10 and ImageNet, respectively), we analyzed the orientation selectivity of individual units. Orientation selectivity is a low-level selectivity widely studied in visual neuroscience, in which, when images of bars with several orientations are presented to the eye, many neurons in the visual cortex respond selectively to a specific orientation. We found that orientation-selective units exist in both lower and higher layers of these DNNs, as in our brain. In particular, units in the lower layers become more orientation-selective as the generalization performance improves during the course of training of the DNNs. Consistently, networks that generalize better are more orientation-selective in the lower layers. We finally reveal that ablating these selective units in the lower layers substantially degrades the generalization performance, at least by disrupting the shift-invariance of the higher layers. These results suggest to the machine-learning community that, contrary to the triviality of units with high-level selectivity, lower-layer units with selectivity for low-level features can be indispensable for generalization, and for neuroscientists, orientation selectivity can play a causally important role in object recognition.

\*\*\*\*\*

Sparse Binary Compression: Towards Distributed Deep Learning with minimal Communication

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, Wojciech Samek

Currently, progressively larger deep neural networks are trained on ever growing data corpora. In result, distributed training schemes are becoming increasingly relevant. A major issue in distributed training is the limited communication bandwidth between contributing nodes or prohibitive communication cost in general.

These challenges become even more pressing, as the number of computation nodes increases.

To mitigate this problem we propose Sparse Binary Compression (SBC), a compression framework that allows for a drastic reduction of communication cost for distributed training. SBC combines existing techniques of communication delay and gradient sparsification with a novel binarization method and optimal weight update encoding to push compression gains to new limits. By doing so, our method also allows us to smoothly trade-off gradient sparsity and temporal sparsity to adapt to the requirements of the learning task.

We use tools from information theory to reason why SBC can achieve the striking compression rates observed in the experiments.

Our experiments show, that SBC can reduce the upstream communication on a variety of convolutional and recurrent neural network architectures by more than four orders of magnitude without significantly harming the convergence speed in terms of forward-backward passes. For instance, we can train ResNet50 on ImageNet in the same number of iterations to the baseline accuracy, using  $\times 3531$  less bits or train it to a  $1\%$  lower accuracy using  $\times 37208$  less bits. In the latter case, the total upstream communication required is cut from 125 terabytes to 3.35 gigabytes for every participating client. Our method also achieves state-of-the-art compression rates in a Federated Learning setting with 400 clients.

\*\*\*\*\*

Large-Scale Study of Curiosity-Driven Learning

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, Alexei A. Efros

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.  $\{\text{without any extrinsic rewards}\}$ , 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/>.

\*\*\*\*\*

On Tighter Generalization Bounds for Deep Neural Networks: CNNs, ResNets, and Beyond

Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, Tuo Zhao

We propose a generalization error bound for a general family of deep neural networks based on the depth and width of the networks, as well as the spectral norm of weight matrices. Through introducing a novel characterization of the Lipschitz properties of neural network family, we achieve a tighter generalization error bound. We further obtain a result that is free of linear dependence on norms for bounded losses. Besides the general deep neural networks, our results can be applied to derive new bounds for several popular architectures, including convolutional neural networks (CNNs), residual networks (ResNets), and hyperspherical networks (SphereNets). When achieving same generalization errors with previous arts, our bounds allow for the choice of much larger parameter spaces of weight matrices, inducing potentially stronger expressive ability for neural networks.

\*\*\*\*\*

Intriguing Properties of Learned Representations

Amartya Sanyal, Varun Kanade, Philip H. Torr

A key feature of neural networks, particularly deep convolutional neural networks, is their ability to learn useful representations from data. The very last layer of a neural network is then simply a linear model trained on these learned representations. Despite their numerous applications in other tasks such as classification, retrieval, clustering etc., a.k.a. transfer learning, not much work has been published that investigates the structure of these representations or indeed whether structure can be imposed on them during the training process.

In this paper, we study the effective dimensionality of the learned representations by models that have proved highly successful for image classification. We focus on ResNet-18, ResNet-50 and VGG-19 and observe that when trained on CIFAR 10 or CIFAR100, the learned representations exhibit a fairly low rank structure.

We propose a modification to the training procedure, which further encourages low rank structure on learned activations. Empirically, we show that this has implications for robustness to adversarial examples and compression.

\*\*\*\*\*

StrokeNet: A Neural Painting Environment

Ningyuan Zheng, Yifan Jiang, Dingjiang Huang

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

d with the emulated environment is able to directly transfer its skills to real-world software.

\*\*\*\*\*

Pushing the bounds of dropout

Gábor Melis, Charles Blundell, Tomáš Kořmiský, Karl Moritz Hermann, Chris Dyer, Phil Blunsom

We show that dropout training is best understood as performing MAP estimation concurrently for a family of conditional models whose objectives are themselves lower bounded by the original dropout objective. This discovery allows us to pick any model from this family after training, which leads to a substantial improvement on regularisation-heavy language modelling. The family includes models that compute a power mean over the sampled dropout masks, and their less stochastic subvariants with tighter and higher lower bounds than the fully stochastic dropout objective. We argue that since the deterministic subvariant's bound is equal to its objective, and the highest amongst these models, the predominant view of it as a good approximation to MC averaging is misleading. Rather, deterministic dropout is the best available approximation to the true objective.

\*\*\*\*\*

Deep processing of structured data

Lukasz Maziarka, Marek Mnieja, Aleksandra Nowak, Jacek Tabor, Lukasz Struski, Przemysław Spurek

We construct a general unified framework for learning representation of structured data, i.e. data which cannot be represented as the fixed-length vectors (e.g. sets, graphs, texts or images of varying sizes). The key factor is played by an intermediate

network called SAN (Set Aggregating Network), which maps a structured object to a fixed length vector in a high dimensional latent space. Our main theoretical

result shows that for sufficiently large dimension of the latent space, SAN is capable of learning a unique representation for every input example. Experiments demonstrate that replacing pooling operation by SAN in convolutional networks leads to better results in classifying images with different sizes. Moreover, its direct

application to text and graph data allows to obtain results close to SOTA, by simpler networks with smaller number of parameters than competitive models.

\*\*\*\*\*

Dynamic Early Terminating of Multiply Accumulate Operations for Saving Computation Cost in Convolutional Neural Networks

Yu-Yi Su, Yung-Chih Chen, Xiang-Xiu Wu, Shih-Chieh Chang

Deep learning has been attracting enormous attention from academia as well as in industry due to its great success in many artificial intelligence applications. As more applications are developed, the need for implementing a complex neural network model on an energy-limited edge device becomes more critical. To this end, this paper proposes a new optimization method to reduce the computation efforts of convolutional neural networks. The method takes advantage of the fact that some convolutional operations are actually wasteful since their outputs are pruned by the following activation or pooling layers. Basically, a convolutional filter conducts a series of multiply-accumulate (MAC) operations. We propose to set a checkpoint in the MAC process to determine whether a filter could terminate early based on the intermediate result. Furthermore, a fine-tuning process is conducted to recover the accuracy drop due to the applied checkpoints. The experimental results show that the proposed method can save approximately 50% MAC operations with less than 1% accuracy drop for CIFAR-10 example model and Network in Network on the CIFAR-10 and CIFAR-100 datasets. Additionally, compared with the state-of-the-art method, the proposed method is more effective on the CIFAR-10 dataset and is competitive on the CIFAR-100 dataset.

\*\*\*\*\*

Laplacian Networks: Bounding Indicator Function Smoothness for Neural Networks R

robustness

Carlos Eduardo Rosar Kos Lassance, Vincent Gripon, Antonio Ortega

For the past few years, Deep Neural Network (DNN) robustness has become a question of paramount importance. As a matter of fact, in sensitive settings misclassification can lead to dramatic consequences. Such misclassifications are likely to occur when facing adversarial attacks, hardware failures or limitations, and imperfect signal acquisition. To address this question, authors have proposed different approaches aiming at increasing the robustness of DNNs, such as adding regularizers or training using noisy examples. In this paper we propose a new regularizer built upon the Laplacian of similarity graphs obtained from the representation of training data at each layer of the DNN architecture. This regularizer penalizes large changes (across consecutive layers in the architecture) in the distance between examples of different classes, and as such enforces smooth variations of the class boundaries. Since it is agnostic to the type of deformations that are expected when predicting with the DNN, the proposed regularizer can be combined with existing ad-hoc methods. We provide theoretical justification for this regularizer and demonstrate its effectiveness to improve robustness of DNNs on classical supervised learning vision datasets.

\*\*\*\*\*

Low Latency Privacy Preserving Inference

Alon Brutzkus, Oren Elisha, Ran Gilad-Bachrach

When applying machine learning to sensitive data one has to balance between accuracy, information leakage, and computational-complexity. Recent studies have shown that Homomorphic Encryption (HE) can be used for protecting against information leakage while applying neural networks. However, this comes with the cost of limiting the kind of neural networks that can be used (and hence the accuracy) and with latency of the order of several minutes even for relatively simple networks. In this study we improve on previous results both in the kind of networks that can be applied and in terms of the latency. Most of the improvement is achieved by novel ways to represent the data to make better use of the capabilities of the encryption scheme.

\*\*\*\*\*

DialogWAE: Multimodal Response Generation with Conditional Wasserstein Auto-Encoder

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, Sunghun Kim

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.

\*\*\*\*\*

RETHINKING SELF-DRIVING : MULTI -TASK KNOWLEDGE FOR BETTER GENERALIZATION AND ACCIDENT EXPLANATION ABILITY

Zhihao LI, Toshiyuki MOTOYOSHI, Kazuma SASAKI, Tetsuya OGATA, Shigeki SUGANO

Current end-to-end deep learning driving models have two problems: (1) Poor generalization ability of unobserved driving environment when diversity of training driving dataset is limited (2) Lack of accident explanation ability when driving

models don't work as expected. To tackle these two problems, rooted on the believe that knowledge of associated easy task is beneficial for addressing difficult

task, we proposed a new driving model which is composed of perception module for see and think and driving module for behave, and trained it with multi-task perception-related basic knowledge and driving knowledge stepwisely. Specifically segmentation map and depth map (pixel level understanding of images) were considered as what & where and how far knowledge for tackling easier driving-related perception problems before generating final control commands for difficult

driving task. The results of experiments demonstrated the effectiveness of multi-

task perception knowledge for better generalization and accident explanation ability.

With our method the average success rate of finishing most difficult navigation

tasks in untrained city of CoRL test surpassed current benchmark method for 15 percent in trained weather and 20 percent in untrained weathers.

\*\*\*\*\*

#### Quaternion Recurrent Neural Networks

Titouan Parcollet, Mirco Ravanelli, Mohamed Morchid, Georges Linarès, Chiheb Trabelsi, Renato De Mori, Yoshua Bengio

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

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.

\*\*\*\*\*

#### Optimization on Multiple Manifolds

Mingyang Yi, Huishuai Zhang, Wei Chen, Zhi-ming Ma, Tie-yan Liu

Optimization on manifold has been widely used in machine learning, to handle optimization problems with constraint. Most previous works focus on the case with a single manifold. However, in practice it is quite common that the optimization problem involves more than one constraints, (each constraint corresponding to one manifold). It is not clear in general how to optimize on multiple manifolds effectively and provably especially when the intersection of multiple manifolds is

not a manifold or cannot be easily calculated. We propose a unified algorithm framework to handle the optimization on multiple manifolds. Specifically, we integrate information from multiple manifolds and move along an ensemble direction by viewing the information from each manifold as a drift and adding them together. We prove the convergence properties of the proposed algorithms. We also apply the algorithms into training neural network with batch normalization layers and achieve preferable empirical results.

\*\*\*\*\*

Invariant-equivariant representation learning for multi-class data

Ilya Feige

Representations learnt through deep neural networks tend to be highly informative, but opaque in terms of what information they learn to encode. We introduce an approach to probabilistic modelling that learns to represent data with two separate deep representations: an invariant representation that encodes the information of the class from which the data belongs, and an equivariant representation that encodes the symmetry transformation defining the particular data point within the class manifold (equivariant in the sense that the representation varies naturally with symmetry transformations). This approach to representation learning is conceptually transparent, easy to implement, and in-principle generally applicable to any data comprised of discrete classes of continuous distributions (e.g. objects in images, topics in language, individuals in behavioural data). We demonstrate qualitatively compelling representation learning and competitive quantitative performance, in both supervised and semi-supervised settings, versus comparable modelling approaches in the literature with little fine tuning.

\*\*\*\*\*

Language Modeling Teaches You More Syntax than Translation Does: Lessons Learned Through Auxiliary Task Analysis

Kelly W. Zhang, Samuel R. Bowman

Recent work using auxiliary prediction task classifiers to investigate the properties of LSTM representations has begun to shed light on why pretrained representations, like ELMo (Peters et al., 2018) and CoVe (McCann et al., 2017), are so beneficial for neural language understanding models. We still, though, do not yet have a clear understanding of how the choice of pretraining objective affects the type of linguistic information that models learn. With this in mind, we compare four objectives - language modeling, translation, skip-thought, and autoencoding - on their ability to induce syntactic and part-of-speech information. We make a fair comparison between the tasks by holding constant the quantity and genre of the training data, as well as the LSTM architecture. We find that representations from language models consistently perform best on our syntactic auxiliary prediction tasks, even when trained on relatively small amounts of data. These results suggest that language modeling may be the best data-rich pretraining task for transfer learning applications requiring syntactic information. We also find that the representations from randomly-initialized, frozen LSTMs perform strikingly well on our syntactic auxiliary tasks, but this effect disappears when the amount of training data for the auxiliary tasks is reduced.

\*\*\*\*\*

Convolutional CRFs for Semantic Segmentation

Marvin Teichmann, Roberto Cipolla

For the challenging semantic image segmentation task the best performing models have traditionally combined the structured modelling capabilities of Conditional Random Fields (CRFs) with the feature extraction power of CNNs. In more recent works however, CRF post-processing has fallen out of favour. We argue that this is mainly due to the slow training and inference speeds of CRFs, as well as the difficulty of learning the internal CRF parameters. To overcome both issues we propose to add the assumption of conditional independence to the framework of fully-connected CRFs. This allows us to reformulate the inference in terms of convolutions, which can be implemented highly efficiently on GPUs. Doing so speeds up inference and training by two orders of magnitude. All parameters of the convolutional CRFs can easily be optimized using backpropagation. Towards the goal of facilitating further CRF research we have made our implementations

publicly available.

\*\*\*\*\*

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

Xiaopeng Li, Zhourong Chen, Leonard K. M. Poon, Nevin L. Zhang

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.

\*\*\*\*\*

## Sentence Encoding with Tree-Constrained Relation Networks

Lei Yu, Cyprien de Masson d'Autume, Chris Dyer, Phil Blunsom, Lingpeng Kong, Wang Ling

The meaning of a sentence is a function of the relations that hold between its words. We instantiate this relational view of semantics in a series of neural models based on variants of relation networks (RNs) which represent a set of objects (for us, words forming a sentence) in terms of representations of pairs of objects. We propose two extensions to the basic RN model for natural language. First, building on the intuition that not all word pairs are equally informative about the meaning of a sentence, we use constraints based on both supervised and unsupervised dependency syntax to control which relations influence the representation. Second, since higher-order relations are poorly captured by a sum of pairwise relations, we use a recurrent extension of RNs to propagate information so as to form representations of higher order relations. Experiments on sentence classification, sentence pair classification, and machine translation reveal that, while basic RNs are only modestly effective for sentence representation, recurrent RNs with latent syntax are a reliably powerful representational device.

\*\*\*\*\*

## ReNeg and Backseat Driver: Learning from demonstration with continuous human feedback

Zoe Papakipos, Jacob Beck, Michael Littman

Reinforcement learning (RL) is a powerful framework for solving problems by exploring and learning from mistakes. However, in the context of autonomous vehicle (AV) control, requiring an agent to make mistakes, or even allowing mistakes, can be quite dangerous and costly in the real world. For this reason, AV RL is generally only viable in simulation. Because these simulations have imperfect representations, particularly with respect to graphics, physics, and human interaction, we find motivation for a framework similar to RL, suitable to the real world.

To this end, we formulate a learning framework that learns from restricted exploration by having a human demonstrator do the exploration. Existing work on learning from demonstration typically either assumes the collected data is performed by an optimal expert, or requires potentially dangerous exploration to find the optimal policy. We propose an alternative framework that learns continuous control from only safe behavior. One of our key insights is that the problem becomes tractable if the feedback score that rates the demonstration applies to the atomic action, as opposed to the entire sequence of actions. We use human experts to collect driving data as well as to label the driving data through a framework we call "Backseat Driver", giving us state-action pairs matched with scalar values representing the score for the action. We call the more general learning framework ReNeg, since it learns a regression from states to actions given negative as well as positive examples. We empirically validate several models in the ReNeg framework, testing on lane-following with limited data. We find that the best solution in this context outperforms behavioral cloning has strong connections



to stochastic policy gradient approaches.

\*\*\*\*\*

#### Generalization and Regularization in DQN

Jesse Farebrother, Marlos C. Machado, Michael Bowling

Deep reinforcement learning (RL) algorithms have shown an impressive ability to learn complex control policies in high-dimensional environments. However, despite the ever-increasing performance on popular benchmarks like the Arcade Learning Environment (ALE), policies learned by deep RL algorithms can struggle to generalize when evaluated in remarkably similar environments. These results are unexpected given the fact that, in supervised learning, deep neural networks often learn robust features that generalize across tasks. In this paper, we study the generalization capabilities of DQN in order to aid in understanding this mismatch between generalization in deep RL and supervised learning methods. We provide evidence suggesting that DQN overspecializes to the domain it is trained on. We then comprehensively evaluate the impact of traditional methods of regularization from supervised learning,  $\ell_2$  and dropout, and of reusing learned representations to improve the generalization capabilities of DQN. We perform this study using different game modes of Atari 2600 games, a recently introduced modification for the ALE which supports slight variations of the Atari 2600 games used for benchmarking in the field. Despite regularization being largely underutilized in deep RL, we show that it can, in fact, help DQN learn more general features. These features can then be reused and fine-tuned on similar tasks, considerably improving the sample efficiency of DQN.

\*\*\*\*\*

#### An Automatic Operation Batching Strategy for the Backward Propagation of Neural Networks Having Dynamic Computation Graphs

Yuchen Qiao, Kenjiro Taura

Organizing the same operations in the computation graph of a neural network into batches is one of the important methods to improve the speed of training deep learning models and applications since it helps to execute operations with the same type in parallel and to make full use of the available hardware resources. This batching task is usually done by the developers manually and it becomes more difficult when the neural networks have dynamic computation graphs because of the input data with varying structures or the dynamic flow control. Several automatic batching strategies were proposed and integrated into some deep learning toolkits so that the programmers don't have to be responsible for this task. These strategies, however, will miss some important opportunities to group the operations in the backward propagation of training neural networks. In this paper, we proposed a strategy which provides more efficient automatic batching and brings benefits to the memory access in the backward propagation. We also test our strategy on a variety of benchmarks with dynamic computation graphs. The result shows that it really brings further improvements in the training speed when our strategy is working with the existing automatic strategies.

\*\*\*\*\*

#### Variational Smoothing in Recurrent Neural Network Language Models

Lingpeng Kong, Gabor Melis, Wang Ling, Lei Yu, Dani Yogatama

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.

\*\*\*\*\*

#### Curiosity-Driven Experience Prioritization via Density Estimation

Rui Zhao, Volker Tresp

In Reinforcement Learning (RL), an agent explores the environment and collects trajectories into the memory buffer for later learning. However, the collected trajectories can easily be imbalanced with respect to the achieved goal states. The problem of learning from imbalanced data is a well-known problem in supervised learning, but has not yet been thoroughly researched in RL. To address this problem, we propose a novel Curiosity-Driven Prioritization (CDP) framework to encourage the agent to over-sample those trajectories that have rare achieved goal states. The CDP framework mimics the human learning process and focuses more on relatively uncommon events. We evaluate our methods using the robotic environment provided by OpenAI Gym. The environment contains six robot manipulation tasks. In our experiments, we combined CDP with Deep Deterministic Policy Gradient (DDPG) with or without Hindsight Experience Replay (HER). The experimental results show that CDP improves both performance and sample-efficiency of reinforcement learning agents, compared to state-of-the-art methods.

\*\*\*\*\*

#### Initialized Equilibrium Propagation for Backprop-Free Training

Peter O'Connor, Efstratios Gavves, Max Welling

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 training 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 feedforward 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.

\*\*\*\*\*

#### Unsupervised Expectation Learning for Multisensory Binding

Pablo Barros, German I. Parisi, Manfred Eppe, Stefan Wermter

Expectation learning is a continuous learning process which uses known multisensory bindings to modulate unisensory perception. When perceiving an event, we have an expectation on what we should see or hear which affects our unisensory perception. Expectation learning is known to enhance the unisensory perception of previously known multisensory events. In this work, we present a novel hybrid deep recurrent model based on audio/visual autoencoders, for unimodal stimulus representation and reconstruction, and a recurrent self-organizing network for multisensory binding of the representations. The model adapts concepts of expectation learning to enhance the unisensory representation based on the learned bindings. We demonstrate that the proposed model is capable of reconstructing signals from one modality by processing input of another modality for 43,500 Youtube videos in the animal subset of the AudioSet corpus. Our experiments also show that when using expectation learning, the proposed model presents state-of-the-art performance in representing and classifying unisensory stimuli.

\*\*\*\*\*

#### Multi-turn Dialogue Response Generation in an Adversarial Learning Framework

Oluwatobi O. Olabiyi, Alan Salimov, Anish Khazane, Erik T. Mueller

We propose an adversarial learning approach to the generation of multi-turn dialogue responses. Our proposed framework, hredGAN, is based on conditional generative adversarial networks (GANs). The GAN's generator is a modified hierarchical recurrent encoder-decoder network (HRED) and the discriminator is a word-level b

idirectional RNN that shares context and word embedding with the generator. During inference, noise samples conditioned on the dialogue history are used to perturb the generator's latent space to generate several possible responses. The final response is the one ranked best by the discriminator. The hredGAN shows major advantages over existing methods: (1) it generalizes better than networks trained using only the log-likelihood criterion, and (2) it generates longer, more informative and more diverse responses with high utterance and topic relevance even with limited training data. This superiority is demonstrated on the Movie triples and Ubuntu dialogue datasets with both the automatic and human evaluations.

\*\*\*\*\*

The loss landscape of overparameterized neural networks

Y. Cooper

We explore some mathematical features of the loss landscape of overparameterized neural networks. A priori one might imagine that the loss function looks like a typical function from  $\mathbb{R}^n$  to  $\mathbb{R}$  - in particular, nonconvex, with discrete global minima. In this paper, we prove that in at least one important way, the loss function of an overparameterized neural network does not look like a typical function. If a neural net has  $n$  parameters and is trained on  $d$  data points, with  $n > d$ , we show that the locus  $\mathcal{M}$  of global minima of  $L$  is usually not discrete, but rather an  $n-d$  dimensional submanifold of  $\mathbb{R}^n$ . In practice, neural nets commonly have orders of magnitude more parameters than data points, so this observation implies that  $\mathcal{M}$  is typically a very high-dimensional subset of  $\mathbb{R}^n$ .

\*\*\*\*\*

Identifying and Controlling Important Neurons in Neural Machine Translation

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, James Glass

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.

\*\*\*\*\*

Switching Linear Dynamics for Variational Bayes Filtering

Philip Becker-Ehmck, Jan Peters, Patrick van der Smagt

System identification of complex and nonlinear systems is a central problem for model predictive control and model-based reinforcement learning. Despite their complexity, such systems can often be approximated well by a set of linear dynamical systems if broken into appropriate subsequences. This mechanism not only helps us find good approximations of dynamics, but also gives us deeper insight into the underlying system. Leveraging Bayesian inference and Variational Autoencoders, we show how to learn a richer and more meaningful state space, e.g. encoding joint constraints and collisions with walls in a maze, from partial and high-dimensional observations. This representation translates into a gain of accuracy of the learned dynamics which we showcase on various simulated tasks.

\*\*\*\*\*

signSGD via Zeroth-Order Oracle

Sijia Liu, Pin-Yu Chen, Xiangyi Chen, Mingyi Hong

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 prop

ose 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

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.

\*\*\*\*\*

#### RelWalk -- A Latent Variable Model Approach to Knowledge Graph Embedding

Danushka Bollegala, Huda Hakami, Yuichi Yoshida, Ken-ichi Kawarabayashi

Knowledge Graph Embedding (KGE) is the task of jointly learning entity and relation embeddings for a given knowledge graph. Existing methods for learning KGEs can be seen as a two-stage process where (a) entities and relations in the knowledge graph are represented using some linear algebraic structures (embeddings), and (b) a scoring function is defined that evaluates the strength of a relation that holds between two entities using the corresponding relation and entity embeddings. Unfortunately, prior proposals for the scoring functions in the first step have been heuristically motivated, and it is unclear as to how the scoring functions in KGEs relate to the generation process of the underlying knowledge graph. To address this issue, we propose a generative account of the KGE learning task. Specifically, given a knowledge graph represented by a set of relational triples  $(h, R, t)$ , where the semantic relation  $R$  holds between the two entities  $h$  (head) and  $t$  (tail), we extend the random walk model (Arora et al., 2016a) of word embeddings to KGE. We derive a theoretical relationship between the joint probability  $p(h, R, t)$  and the embeddings of  $h$ ,  $R$  and  $t$ . Moreover, we show that marginal loss minimisation, a popular objective used by much prior work in KGE, follows naturally from the log-likelihood ratio maximisation under the probabilities estimated from the KGEs according to our theoretical relationship. We propose a learning objective motivated by the theoretical analysis to learn KGEs from a given knowledge graph. The KGEs learnt by our proposed method obtain state-of-the-art performance on FB15K237 and WN18RR benchmark datasets, providing empirical evidence in support of the theory.

\*\*\*\*\*

#### DynCNN: An Effective Dynamic Architecture on Convolutional Neural Network for Surveillance Videos

De-Qin Gao, Ping-Chen Tsai, Shanq-Jang Ruan

The large-scale surveillance video analysis becomes important as the development of intelligent city. The heavy computation resources necessary for state-of-the-art deep learning model makes the real-time processing hard to be implemented. This paper exploits the characteristic of high scene similarity generally existing in surveillance videos and proposes dynamic convolution reusing the previous

feature map to reduce the computation amount. We tested the proposed method on 45 surveillance videos with various scenes. The experimental results show that dynamic convolution can reduce up to 75.7% of FLOPs while preserving the precision within 0.7% mAP. Furthermore, the dynamic convolution can enhance the processing time up to 2.2 times.

\*\*\*\*\*

Globally Soft Filter Pruning For Efficient Convolutional Neural Networks

Ke Xu, Xiaoyun Wang, Qun Jia, Jianjing An, Dong Wang

This paper propose a cumulative saliency based Globally Soft Filter Pruning (GSFP) scheme to prune redundant filters of Convolutional Neural Networks (CNNs). Specifically, the GSFP adopts a robust pruning method, which measures the global redundancy of the filter in the whole model by using the soft pruning strategy. In addition, in the model recovery process after pruning, we use the cumulative saliency strategy to improve the accuracy of pruning. GSFP has two advantages over previous works: (1) More accurate pruning guidance. For a pre-trained CNN model, the saliency of the filter varies with different input data. Therefore, accumulating the saliency of the filter over the entire data set can provide more accurate guidance for pruning. On the other hand, pruning from a global perspective is more accurate than local pruning. (2) More robust pruning strategy. We propose a reasonable normalization formula to prevent certain layers of filters in the network from being completely clipped due to excessive pruning rate.

\*\*\*\*\*

PRUNING IN TRAINING: LEARNING AND RANKING SPARSE CONNECTIONS IN DEEP CONVOLUTIONAL NETWORKS

Yanwei Fu, Shun Zhang, Donghao Li, Xinwei Sun, Xiangyang Xue, Yuan Yao

This paper proposes a Pruning in Training (PiT) framework of learning to reduce the parameter size of networks. Different from existing works, our PiT framework employs the sparse penalties to train networks and thus help rank the importance of weights and filters. Our PiT algorithms can directly prune the network without any fine-tuning. The pruned networks can still achieve comparable performance to the original networks. In particular, we introduce the (Group) Lasso-type Penalty (L-P / GL-P), and (Group) Split LBI Penalty (S-P / GS-P) to regularize the networks, and a pruning strategy proposed is used in help prune the network. We conduct the extensive experiments on MNIST, Cifar-10, and miniImageNet. The results validate the efficacy of our proposed methods. Remarkably, on MNIST dataset, our PiT framework can save 17.5% parameter size of LeNet-5, which achieves the 98.47% recognition accuracy.

\*\*\*\*\*

Learning to Remember More with Less Memorization

Hung Le, Truyen Tran, Svetha Venkatesh

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.

\*\*\*\*\*

Log Hyperbolic Cosine Loss Improves Variational Auto-Encoder

Pengfei Chen, Guangyong Chen, Shengyu Zhang

In Variational Auto-Encoder (VAE), the default choice of reconstruction loss function between the decoded sample and the input is the squared  $L_2$ . We propose

to replace it with the log hyperbolic cosine (log-cosh) loss, which behaves as  $L_2$  at small values and as  $L_1$  at large values, and differentiable everywhere. Compared with  $L_2$ , the log-cosh loss improves the reconstruction without damaging the latent space optimization, thus automatically keeping a balance between the reconstruction and the generation. Extensive experiments on MNIST and CelebA datasets show that the log-cosh reconstruction loss significantly improves the performance of VAE and its variants in output quality, measured by sharpness and FID score. In addition, the gradient of the log-cosh is a simple tanh function, which makes the implementation of gradient descent as simple as adding one sentence in coding.

\*\*\*\*\*

#### Convolutional Neural Networks combined with Runge-Kutta Methods

Mai Zhu, Bo Chang, Chong Fu

A convolutional neural network for image classification can be constructed mathematically since it can be regarded as a multi-period dynamical system. In this paper, a novel approach is proposed to construct network models from the dynamical systems view. Since a pre-activation residual network can be deemed an approximation of a time-dependent dynamical system using the forward Euler method, higher order Runge-Kutta methods (RK methods) can be utilized to build network models in order to achieve higher accuracy. The model constructed in such a way is referred to as the Runge-Kutta Convolutional Neural Network (RKNet). RK methods also provide an interpretation of Dense Convolutional Networks (DenseNets) and Convolutional Neural Networks with Alternately Updated Clique (CliqueNets) from the dynamical systems view. The proposed methods are evaluated on benchmark datasets: CIFAR-10/100, SVHN and ImageNet. The experimental results are consistent with the theoretical properties of RK methods and support the dynamical systems interpretation. Moreover, the experimental results show that the RKNets are superior to the state-of-the-art network models on CIFAR-10 and on par on CIFAR-100, SVHN and ImageNet.

\*\*\*\*\*

#### Discrete Structural Planning for Generating Diverse Translations

Raphael Shu, Hideki Nakayama

Planning is important for humans when producing complex languages, which is a missing part in current language generation models. In this work, we add a planning phase in neural machine translation to control the global sentence structure ahead of translation. Our approach learns discrete structural representations to encode syntactic information of target sentences. During translation, we can either let beam search to choose the structural codes automatically or specify the codes manually. The word generation is then conditioned on the selected discrete codes. Experiments show that the translation performance remains intact by learning the codes to capture pure structural variations. Through structural planning, we are able to control the global sentence structure by manipulating the codes. By evaluating with a proposed structural diversity metric, we found that the sentences sampled using different codes have much higher diversity scores. In qualitative analysis, we demonstrate that the sampled paraphrase translations have drastically different structures.

\*\*\*\*\*

#### Mapping the hyponymy relation of wordnet onto vector Spaces

Jean-Philippe Bernardy, Aleksandre Maskharashvili

In this paper, we investigate mapping the hyponymy relation of wordnet to feature vectors.

We aim to model lexical knowledge in such a way that it can be used as input in generic machine-learning models, such as phrase entailment predictors.

We propose two models. The first one leverages an existing mapping of words to feature vectors (fasttext), and attempts to classify

such vectors as within or outside of each class. The second model is fully supervised,

using solely wordnet as a ground truth. It maps each concept to an interval or a disjunction thereof.

On the first model, we approach, but not quite attain state of the art performance. The second model can achieve near-perfect accuracy.

\*\*\*\*\*

On the Selection of Initialization and Activation Function for Deep Neural Networks

Soufiane Hayou, Arnaud Doucet, Judith Rousseau

The weight initialization and the activation function of deep neural networks have a crucial impact on the performance of the training procedure. An inappropriate selection can lead to the loss of information of the input during forward propagation and the exponential vanishing/exploding of gradients during back-propagation. Understanding the theoretical properties of untrained random networks is key to identifying which deep networks may be trained successfully as recently demonstrated by Schoenholz et al. (2017) who showed that for deep feedforward neural networks only a specific choice of hyperparameters known as the 'edge of chaos' can lead to good performance.

We complete this analysis by providing quantitative results showing that, for a class of ReLU-like activation functions, the information propagates indeed deeper for an initialization at the edge of chaos. By further extending this analysis, we identify a class of activation functions that improve the information propagation over ReLU-like functions. This class includes the Swish activation,  $\phi_{\text{swish}}(x) = x \cdot \text{sigmoid}(x)$ , used in Hendrycks & Gimpel (2016), Elfving et al. (2017) and Ramachandran et al. (2017). This provides a theoretical grounding for the excellent empirical performance of  $\phi_{\text{swish}}$  observed in these contributions. We complement those previous results by illustrating the benefit of using a random initialization on the edge of chaos in this context.

\*\*\*\*\*

An Information-Theoretic Metric of Transferability for Task Transfer Learning

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Amir R. Zamir, Leonidas J. Guibas

An important question in task transfer learning is to determine task transferability, i.e. given a common input domain, estimating to what extent representations learned from a source task can help in learning a target task. Typically, transferability is either measured experimentally or inferred through task relatedness, which is often defined without a clear operational meaning. In this paper, we present a novel metric, H-score, an easily-computable evaluation function that estimates the performance of transferred representations from one task to another in classification problems. Inspired by a principled information theoretic approach, H-score has a direct connection to the asymptotic error probability of the decision function based on the transferred feature. This formulation of transferability can further be used to select a suitable set of source tasks in task transfer learning problems or to devise efficient transfer learning policies. Experiments using both synthetic and real image data show that not only our formulation of transferability is meaningful in practice, but also it can generalize to inference problems beyond classification, such as recognition tasks for 3D indoor-scene understanding.

\*\*\*\*\*

Unsupervised Emergence of Spatial Structure from Sensorimotor Prediction

Alban Laflaquière, Michael Garcia Ortiz

Despite its omnipresence in robotics application, the nature of spatial knowledge and the mechanisms that underlie its emergence in autonomous agents are still poorly understood. Recent theoretical work suggests that the concept of space can be grounded by capturing invariants induced by the structure of space in an agent's raw sensorimotor experience. Moreover, it is hypothesized that capturing these invariants is beneficial for a naive agent trying to predict its sensorimotor experience. Under certain exploratory conditions, spatial representations should thus emerge as a byproduct of learning to predict.

We propose a simple sensorimotor predictive scheme, apply it to different agents and types of exploration, and evaluate the pertinence of this hypothesis. We show that a naive agent can capture the topology and metric regularity of its spatial configuration without any a priori knowledge, nor extraneous supervision.

\*\*\*\*\*

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

Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, Dmitry Vetrov

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.

\*\*\*\*\*

## Exploring and Enhancing the Transferability of Adversarial Examples

Lei Wu, Zhanxing Zhu, Cheng Tai

State-of-the-art deep neural networks are vulnerable to adversarial examples, formed by applying small but malicious perturbations to the original inputs. Moreover, the perturbations can \textit{transfer across models}: adversarial examples generated for a specific model will often mislead other unseen models. Consequently the adversary can leverage it to attack deployed systems without any query, which severely hinders the application of deep learning, especially in the safety-critical areas. In this work, we empirically study how two classes of factors those might influence the transferability of adversarial examples. One is about model-specific factors, including network architecture, model capacity and test accuracy. The other is the local smoothness of loss surface for constructing adversarial examples. Inspired by these understandings on the transferability of adversarial examples, we then propose a simple but effective strategy to enhance the transferability, whose effectiveness is confirmed by a variety of experiments on both CIFAR-10 and ImageNet datasets.

\*\*\*\*\*

## Deterministic Variational Inference for Robust Bayesian Neural Networks

Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E. Turner, José Miguel Hernández-Lobato, Alexander L. Gaunt

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.

\*\*\*\*\*

## Learning to Separate Domains in Generalized Zero-Shot and Open Set Learning: a probabilistic perspective

Hanze Dong, Yanwei Fu, Leonid Sigal, Sungju Hwang, Yu-Gang Jiang, Xiangyang Xue

This paper studies the problem of domain division which aims to segment instances drawn from different probabilistic distributions. This problem exists in many



previous recognition tasks, such as Open Set Learning (OSL) and Generalized Zero-Shot Learning (G-ZSL), where the testing instances come from either seen or unseen/novel classes with different probabilistic distributions. Previous works only calibrate the confident prediction of classifiers of seen classes (WSVM Scheirer et al. (2014)) or taking unseen classes as outliers Socher et al. (2013). In contrast, this paper proposes a probabilistic way of directly estimating and fine-tuning the decision boundary between seen and unseen classes. In particular, we propose a domain division algorithm to split the testing instances into known, unknown and uncertain domains, and then conduct recognition tasks in each domain. Two statistical tools, namely, bootstrapping and KolmogorovSmirnov (K-S) Test, for the first time, are introduced to uncover and fine-tune the decision boundary of each domain. Critically, the uncertain domain is newly introduced in our framework to adopt those instances whose domain labels cannot be predicted confidently. Extensive experiments demonstrate that our approach achieved the state-of-the-art performance on OSL and G-ZSL benchmarks.

\*\*\*\*\*

Purchase as Reward : Session-based Recommendation by Imagination Reconstruction  
Qibing Li,Xiaolin Zheng

One of the key challenges of session-based recommender systems is to enhance users' purchase intentions. In this paper, we formulate the sequential interactions between user sessions and a recommender agent as a Markov Decision Process (MDP). In practice, the purchase reward is delayed and sparse, and may be buried by clicks, making it an impoverished signal for policy learning. Inspired by the prediction error minimization (PEM) and embodied cognition, we propose a simple architecture to augment reward, namely Imagination Reconstruction Network (IRN). Specifically, IRN enables the agent to explore its environment and learn predictive representations via three key components. The imagination core generates predicted trajectories, i.e., imagined items that users may purchase. The trajectory manager controls the granularity of imagined trajectories using the planning strategies, which balances the long-term rewards and short-term rewards. To optimize the action policy, the imagination-augmented executor minimizes the intrinsic imagination error of simulated trajectories by self-supervised reconstruction, while maximizing the extrinsic reward using model-free algorithms. Empirically, IRN promotes quicker adaptation to user interest, and shows improved robustness to the cold-start scenario and ultimately higher purchase performance compared to several baselines. Somewhat surprisingly, IRN using only the purchase reward achieves excellent next-click prediction performance, demonstrating that the agent can "guess what you like" via internal planning.

\*\*\*\*\*

Directional Analysis of Stochastic Gradient Descent via von Mises-Fisher Distributions in Deep Learning

Cheolhyun Lee,Kyunghyun Cho,Wanmo Kang

Although stochastic gradient descent (SGD) is a driving force behind the recent success of deep learning, our understanding of its dynamics in a high-dimensional parameter space is limited. In recent years, some researchers have used the stochasticity of minibatch gradients, or the signal-to-noise ratio, to better characterize the learning dynamics of SGD. Inspired from these work, we here analyze SGD from a geometrical perspective by inspecting the stochasticity of the norms and directions of minibatch gradients. We propose a model of the directional concentration for minibatch gradients through von Mises-Fisher (VMF) distribution, and show that the directional uniformity of minibatch gradients increases over the course of SGD. We empirically verify our result using deep convolutional networks and observe a higher correlation between the gradient stochasticity and the proposed directional uniformity than that against the gradient norm stochasticity, suggesting that the directional statistics of minibatch gradients is a major factor behind SGD.

\*\*\*\*\*

Holographic and other Point Set Distances for Machine Learning

Lukas Balles,Thomas Fischbacher

We introduce an analytic distance function for moderately sized point sets of kn

own cardinality that is shown to have very desirable properties, both as a loss function as well as a regularizer for machine learning applications. We compare our novel construction to other point set distance functions and show proof of concept experiments for training neural networks end-to-end on point set prediction tasks such as object detection.

\*\*\*\*\*

#### Conditional Network Embeddings

Bo Kang, Jeffrey Lijffijt, Tijl De Bie

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 **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 Leoniak, Igor Sieradzki, Igor Podolak

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.

\*\*\*\*\*

#### Volumetric Convolution: Automatic Representation Learning in Unit Ball

Sameera Ramasinghe, Salman Khan, Nick Barnes

Convolution is an efficient technique to obtain abstract feature representations using hierarchical layers in deep networks. Although performing convolution in Euclidean geometries is fairly straightforward, its extension to other topological spaces---such as a sphere  $S^2$  or a unit ball  $B^3$ ---entails unique challenges.

In this work, we propose a novel "volumetric convolution" operation that can effectively convolve arbitrary functions in  $B^3$ . We develop a theoretical framework for "volumetric convolution" based on Zernike polynomials and efficiently implement it as a differentiable and an easily pluggable layer for deep networks. F

urthermore, our formulation leads to derivation of a novel formula to measure the symmetry of a function in  $B^3$  around an arbitrary axis, that is useful in 3D shape analysis tasks. We demonstrate the efficacy of proposed volumetric convolution operation on a possible use-case i.e., 3D object recognition task.

\*\*\*\*\*

Domain Generalization via Invariant Representation under Domain-Class Dependency  
Kei Akuzawa, Yusuke Iwasawa, Yutaka Matsuo

Learning domain-invariant representation is a dominant approach for domain generalization, where we need to build a classifier that is robust toward domain shifts induced by change of users, acoustic or lighting conditions, etc. However, prior domain-invariance-based methods overlooked the underlying dependency of classes (target variable) on source domains during optimization, which causes the trade-off between classification accuracy and domain-invariance, and often interferes with the domain generalization performance. This study first provides the notion of domain generalization under domain-class dependency and elaborates on the importance of considering the dependency by expanding the analysis of Xie et al. (2017). We then propose a method, invariant feature learning under optimal classifier constraints (IFLOC), which explicitly considers the dependency and maintains accuracy while improving domain-invariance. Specifically, the proposed method regularizes the representation so that it has as much domain information as the class labels, unlike prior methods that remove all domain information. Empirical validations show the superior performance of IFLOC to baseline methods, supporting the importance of the domain-class dependency in domain generalization and the efficacy of the proposed method for overcoming the issue.

\*\*\*\*\*

Semi-supervised Learning with Multi-Domain Sentiment Word Embeddings  
Ran Tian, Yash Agrawal, Kento Watanabe, Hiroya Takamura

Word embeddings are known to boost performance of many NLP tasks such as text classification, meanwhile they can be enhanced by labels at the document level to capture nuanced meaning such as sentiment and topic. Can one combine these two research directions to benefit from both? In this paper, we propose to jointly train a text classifier with a label-enhanced and domain-aware word embedding model, using an unlabeled corpus and only a few labeled data from non-target domains. The embeddings are trained on the unlabeled corpus and enhanced by pseudo labels coming from the classifier, and at the same time are used by the classifier as input and training signals. We formalize this symbiotic cycle in a variational Bayes framework, and show that our method improves both the embeddings and the text classifier, outperforming state-of-the-art domain adaptation and semi-supervised learning techniques. We conduct detailed ablative tests to reveal gains from important components of our approach. The source code and experiment data will be publicly released.

\*\*\*\*\*

Fast adversarial training for semi-supervised learning

Dongha Kim, Yongchan Choi, Jae-Joon Han, Changkyu Choi, Yongdai Kim

In semi-supervised learning, Bad GAN approach is one of the most attractive method due to the intuitional simplicity and powerful performances. Bad GAN learns a classifier with bad samples distributed on complement of the support of the input data. But Bad GAN needs additional architectures, a generator and a density estimation model, which involves huge computation and memory consumption cost. VAT is another good semi-supervised learning algorithm, which utilizes unlabeled data to improve the invariance of the classifier with respect to perturbation of inputs. In this study, we propose a new method by combining the ideas of Bad GAN and VAT. The proposed method generates bad samples of high-quality by use of the adversarial training used in VAT. We give theoretical explanations why the adversarial training is good at both generating bad samples and semi-supervised learning. An advantage of the proposed method is to achieve the competitive performances with much fewer computations. We demonstrate advantages our method by various experiments with well known benchmark image datasets.

\*\*\*\*\*

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, A  ron van den Oord, Oriol Vinyals, Nando de Freitas

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.

\*\*\*\*\*

Object detection deep learning networks for Optical Character Recognition

Christopher Bourez, Aurelien Coquard

In this article, we show how we applied a simple approach coming from deep learning networks for object detection to the task of optical character recognition in order to build image features tailored for documents. In contrast to scene text reading in natural images using networks pretrained on ImageNet, our document reading is performed with small networks inspired by MNIST digit recognition challenge, at a small computational budget and a small stride. The object detection modern frameworks allow a direct end-to-end training, with no other algorithm than the deep learning and the non-max-suppression algorithm to filter the duplicate predictions. The trained weights can be used for higher level models, such as, for example, document classification, or document segmentation.

\*\*\*\*\*

Maximal Divergence Sequential Autoencoder for Binary Software Vulnerability Detection

Tue Le, Tuan Nguyen, Trung Le, Dinh Phung, Paul Montague, Olivier De Vel, Lizhen Qu

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

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

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>.

\*\*\*\*\*

Remember and Forget for Experience Replay

Guido Novati, Petros Koumoutsakos

Experience replay (ER) is crucial for attaining high data-efficiency in off-policy deep reinforcement learning (RL). ER entails the recall of experiences obtained in past iterations to compute gradient estimates for the current policy. However, the accuracy of such updates may deteriorate when the policy diverges from past behaviors, possibly undermining the effectiveness of ER. Previous off-policy RL algorithms mitigated this issue by tuning their hyper-parameters in order to abate policy changes. We propose ReF-ER, a method for active management of experiences in the Replay Memory (RM). ReF-ER forgets experiences that would be too unlikely with the current policy and constrains policy changes within a trust region of the behaviors in the RM. We couple ReF-ER with Q-learning, deterministic policy gradient and off-policy gradient methods to show that ReF-ER reliably improves the performance of continuous-action off-policy RL. We complement ReF-ER with a novel off-policy actor-critic algorithm (RACER) for continuous-action control. RACER employs a computationally efficient closed-form approximation of the action values and is shown to be highly competitive with state-of-the-art algorithms on benchmark problems, while being robust to large hyper-parameter variations.

\*\*\*\*\*

Interactive Agent Modeling by Learning to Probe

Tianmin Shu, Caiming Xiong, Ying Nian Wu, Song-Chun Zhu

The ability of modeling the other agents, such as understanding their intentions and skills, is essential to an agent's interactions with other agents. Conventional agent modeling relies on passive observation from demonstrations. In this work, we propose an interactive agent modeling scheme enabled by encouraging an agent to learn to probe. In particular, the probing agent (i.e. a learner) learns to interact with the environment and with a target agent (i.e., a demonstrator) to maximize the change in the observed behaviors of that agent. Through probing, rich behaviors can be observed and are used for enhancing the agent modeling to learn a more accurate mind model of the target agent. Our framework consists of two learning processes: i) imitation learning for an approximated agent model and ii) pure curiosity-driven reinforcement learning for an efficient probing policy to discover new behaviors that otherwise can not be observed. We have validated our approach in four different tasks. The experimental results suggest that

the agent model learned by our approach i) generalizes better in novel scenarios than the ones learned by passive observation, random probing, and other curiosity-driven approaches do, and ii) can be used for enhancing performance in multiple applications including distilling optimal planning to a policy net, collaboration, and competition. A video demo is available at [https://www.dropbox.com/s/8mz6rd3349tso67/Probing\\_Demo.mov?dl=0](https://www.dropbox.com/s/8mz6rd3349tso67/Probing_Demo.mov?dl=0)

\*\*\*\*\*

An Adversarial Learning Framework for a Persona-based Multi-turn Dialogue Model  
Oluwatobi O. Olabiyi, Anish Khazane, Alan Salimov, Erik T. Mueller

In this paper, we extend the persona-based sequence-to-sequence (Seq2Seq) neural network conversation model to a multi-turn dialogue scenario by modifying the state-of-the-art hredGAN architecture to simultaneously capture utterance attributes such as speaker identity, dialogue topic, speaker sentiments and so on. The proposed system, phredGAN has a persona-based HRED generator (PHRED) and a conditional discriminator. We also explore two approaches to accomplish the conditional discriminator: (1) \$phredGAN\_a\$, a system that passes the attribute representation as an additional input into a traditional adversarial discriminator, and (2) \$phredGAN\_d\$, a dual discriminator system which in addition to the adversarial discriminator, collaboratively predicts the attribute(s) that generated the input utterance. To demonstrate the superior performance of phredGAN over the persona SeqSeq model, we experiment with two conversational datasets, the Ubuntu Dialogue Corpus (UDC) and TV series transcripts from the Big Bang Theory and Friends. Performance comparison is made with respect to a variety of quantitative measures as well as crowd-sourced human evaluation. We also explore the trade-offs from using either variant of \$phredGAN\$ on datasets with many but weak attribute modalities (such as with Big Bang Theory and Friends) and ones with few but strong attribute modalities (customer-agent interactions in Ubuntu dataset).

\*\*\*\*\*

Training generative latent models by variational f-divergence minimization  
Mingtian Zhang, Thomas Bird, Raza Habib, Tianlin Xu, David Barber

Probabilistic models are often trained by maximum likelihood, which corresponds to minimizing a specific form of f-divergence between the model and data distribution. We derive an upper bound that holds for all f-divergences, showing the intuitive result that the divergence between two joint distributions is at least as great as the divergence between their corresponding marginals. Additionally, the f-divergence is not formally defined when two distributions have different supports. We thus propose a noisy version of f-divergence which is well defined in such situations. We demonstrate how the bound and the new version of f-divergence can be readily used to train complex probabilistic generative models of data and that the fitted model can depend significantly on the particular divergence used.

\*\*\*\*\*

Context-adaptive Entropy Model for End-to-end Optimized Image Compression  
Jooyoung Lee, Seunghyun Cho, Seung-Kwon Beack

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](https://github.com/JooyoungLeeETRI/CA_Entropy_Model).

\*\*\*\*\*

DecayNet: A Study on the Cell States of Long Short Term Memories

Nicholas I.H. Kuo, Mehrtash T. Harandi, Hanna Suominen, Nicolas Fourrier, Christian Walder, Gabriela Ferraro

It is unclear whether the extensively applied long-short term memory (LSTM) is an optimised architecture for recurrent neural networks. Its complicated design makes the network hard to analyse and non-immediately clear for its utilities in real-world data. This paper studies LSTMs as systems of difference equations, and takes a theoretical mathematical approach to study consecutive transitions in network variables. Our study shows that the cell state propagation is predominantly controlled by the forget gate. Hence, we introduce DecayNets, LSTMs with monotonically decreasing forget gates, to calibrate cell state dynamics. With recurrent batch normalisation, DecayNet outperforms the previous state of the art for permuted sequential MNIST. The Decay mechanism is also beneficial for LSTM-based optimisers, and decrease optimisee neural network losses more rapidly.

Edit status: Revised paper.

\*\*\*\*\*

#### Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks

Patrick Schwab, Lorenz Linhardt, Walter Karlen

Learning representations for counterfactual inference from observational data is of high practical relevance for many domains, such as healthcare, public policy and economics. Counterfactual inference enables one to answer "What if...?" questions, such as "What would be the outcome if we gave this patient treatment  $t_1$ ?" However, current methods for training neural networks for counterfactual inference on observational data are either overly complex, limited to settings with only two available treatment options, or both. Here, we present Perfect Match (PM), a method for training neural networks for counterfactual inference that is easy to implement, compatible with any architecture, does not add computational complexity or hyperparameters, and extends to any number of treatments. PM is based on the idea of augmenting samples within a minibatch with their propensity-matched nearest neighbours. Our experiments demonstrate that PM outperforms a number of more complex state-of-the-art methods in inferring counterfactual outcomes across several real-world and semi-synthetic datasets.

\*\*\*\*\*

#### Multi-Task Learning for Semantic Parsing with Cross-Domain Sketch

Huan Wang, Yuxiang Hu, Li Dong, Feijun Jiang, Zaiqing Nie

Semantic parsing which maps a natural language sentence into a formal machine-readable representation of its meaning, is highly constrained by the limited annotated training data. Inspired by the idea of coarse-to-fine, we propose a general-to-detailed neural network (GDNN) by incorporating cross-domain sketch (CDS) among utterances and their logic forms. For utterances in different domains, the General Network will extract CDS using an encoder-decoder model in a multi-task learning setup. Then for some utterances in a specific domain, the Detailed Network will generate the detailed target parts using sequence-to-sequence architecture with advanced attention to both utterance and generated CDS. Our experiments show that compared to direct multi-task learning, CDS has improved the performance in semantic parsing task which converts users' requests into meaning representation language (MRL). We also use experiments to illustrate that CDS works by adding some constraints to the target decoding process, which further proves the effectiveness and rationality of CDS.

\*\*\*\*\*

#### Analysis of Quantized Models

Lu Hou, Ruiliang Zhang, James T. Kwok

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.

\*\*\*\*\*

#### Strength in Numbers: Trading-off Robustness and Computation via Adversarially-Trained Ensembles

Edward Grefenstette, Robert Stanforth, Brendan O'Donoghue, Jonathan Uesato, Grzegorz Swirszcz, Pushmeet Kohli

While deep learning has led to remarkable results on a number of challenging problems, researchers have discovered a vulnerability of neural networks in adversarial settings, where small but carefully chosen perturbations to the input can make the models produce extremely inaccurate outputs. This makes these models particularly unsuitable for safety-critical application domains (e.g. self-driving cars) where robustness is extremely important. Recent work has shown that augmenting training with adversarially generated data provides some degree of robustness against test-time attacks. In this paper we investigate how this approach scales as we increase the computational budget given to the defender. We show that increasing the number of parameters in adversarially-trained models increases their robustness, and in particular that ensembling smaller models while adversarially training the entire ensemble as a single model is a more efficient way of spending said budget than simply using a larger single model. Crucially, we show that it is the adversarial training of the ensemble, rather than the ensembling of adversarially trained models, which provides robustness.

\*\*\*\*\*

#### Discrete flow posteriors for variational inference in discrete dynamical systems

Laurence Aitchison, Vincent Adam, Srinivas C. Turaga

Each training step for a variational autoencoder (VAE) requires us to sample from the approximate posterior, so we usually choose simple (e.g. factorised) approximate posteriors in which sampling is an efficient computation that fully exploits its GPU parallelism. However, such simple approximate posteriors are often inefficient, as they eliminate statistical dependencies in the posterior. While it is possible to use normalizing flow approximate posteriors for continuous latents, there is nothing analogous for discrete latents. The most natural approach to model discrete dependencies is an autoregressive distribution, but sampling from such distributions is inherently sequential and thus slow. We develop a fast, parallel sampling procedure for autoregressive distributions based on fixed-point iterations which enables efficient and accurate variational inference in discrete state-space models. To optimize the variational bound, we considered two ways to evaluate probabilities: inserting the relaxed samples directly into the pmf for the discrete distribution, or converting to continuous logistic latent variables and interpreting the K-step fixed-point iterations as a normalizing flow. We found that converting to continuous latent variables gave considerable additional scope for mismatch between the true and approximate posteriors, which resulted in biased inferences, we thus used the former approach. We tested our approach on the neuroscience problem of inferring discrete spiking activity from noisy calcium-imaging data, and found that it gave accurate connectivity estimates in an order of magnitude less time.

\*\*\*\*\*

#### Generating Multiple Objects at Spatially Distinct Locations

Tobias Hinz, Stefan Heinrich, Stefan Wermter

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



ge 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.

\*\*\*\*\*

## Stacking for Transfer Learning

Peng Yuankai

In machine learning tasks, overfitting frequently crops up when the number of samples of target domain is insufficient, for the generalization ability of the classifier is poor in this circumstance. To solve this problem, transfer learning utilizes the knowledge of similar domains to improve the robustness of the learner. The main idea of existing transfer learning algorithms is to reduce the difference between domains by sample selection or domain adaptation. However, no matter what transfer learning algorithm we use, the difference always exists and the hybrid training of source and target data leads to reducing fitting capability of the learner on target domain. Moreover, when the relatedness between domains is too low, negative transfer is more likely to occur. To tackle the problem, we proposed a two-phase transfer learning architecture based on ensemble learning, which uses the existing transfer learning algorithms to train the weak learners in the first stage, and uses the predictions of target data to train the final learner in the second stage. Under this architecture, the fitting capability and generalization capability can be guaranteed at the same time. We evaluated the proposed method on public datasets, which demonstrates the effectiveness and robustness of our proposed method.

\*\*\*\*\*

## Difference-Seeking Generative Adversarial Network

Yi-Lin Sung, Sung-Hsien Hsieh, Soo-Chang Pei, Chun-Shien Lu

We propose a novel algorithm, Difference-Seeking Generative Adversarial Network (DSGAN), developed from traditional GAN. DSGAN considers the scenario that the training samples of target distribution,  $p_t$ , are difficult to collect.

Suppose there are two distributions  $p_{\bar{d}}$  and  $p_d$  such that the density of the target distribution can be the differences between the densities of  $p_{\bar{d}}$  and  $p_d$ . We show how to learn the target distribution  $p_t$  only via samples from  $p_d$  and  $p_{\bar{d}}$  (relatively easy to obtain).

DSGAN has the flexibility to produce samples from various target distributions (e.g. the out-of-distribution). Two key applications, semi-supervised learning and adversarial training, are taken as examples to validate the effectiveness of DSGAN. We also provide theoretical analyses about the convergence of DSGAN.

\*\*\*\*\*

## ANYTIME MINIBATCH: EXPLOITING STRAGGLERS IN ONLINE DISTRIBUTED OPTIMIZATION

Nuwan Ferdinand, Haider Al-Lawati, Stark Draper, Matthew Nokleby

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.

\*\*\*\*\*  
What a difference a pixel makes: An empirical examination of features used by CNNs for categorisation

Gaurav Malhotra, Jeffrey Bowers

Convolutional neural networks (CNNs) were inspired by human vision and, in some settings, achieve a performance comparable to human object recognition. This has lead to the speculation that both systems use similar mechanisms to perform recognition. In this study, we conducted a series of simulations that indicate that there is a fundamental difference between human vision and CNNs: while object recognition in humans relies on analysing shape, CNNs do not have such a shape-bias. We teased apart the type of features selected by the model by modifying the CIFAR-10 dataset so that, in addition to containing objects with shape, the images concurrently contained non-shape features, such as a noise-like mask. When trained on these modified set of images, the model did not show any bias towards selecting shapes as features. Instead it relied on whichever feature allowed it to perform the best prediction -- even when this feature was a noise-like mask or a single predictive pixel amongst 50176 pixels. We also found that regularisation methods, such as batch normalisation or Dropout, did not change this behaviour and neither did past or concurrent experience with images from other datasets.

\*\*\*\*\*  
Computing committor functions for the study of rare events using deep learning with importance sampling

Qianxiao Li, Bo Lin, Weiqing Ren

The committor function is a central object of study in understanding transitions between metastable states in complex systems. However, computing the committor function for realistic systems at low temperatures is a challenging task, due to the curse of dimensionality and the scarcity of transition data. In this paper, we introduce a computational approach that overcomes these issues and achieves good performance on complex benchmark problems with rough energy landscapes. The new approach combines deep learning, importance sampling and feature engineering techniques. This establishes an alternative practical method for studying rare transition events among metastable states of complex, high dimensional systems.

\*\*\*\*\*  
S-System, Geometry, Learning, and Optimization: A Theory of Neural Networks

Shuai Li, Kui Jia

We present a formal measure-theoretical theory of neural networks (NN) built on {\it probability coupling theory}. Particularly, we present an algorithm framework, Hierarchical Measure Group and Approximate System (HMGAS), nicknamed S-System, of which NNs are special cases. In addition to many other results, the framework enables us to prove that 1) NNs implement {\it renormalization group (RG)} using information geometry, which points out that the large scale property to renormalize is dual Bregman divergence and completes the analog between NNs and RG; 2) and under a set of {\it realistic} boundedness and diversity conditions, for {\it large size nonlinear deep} NNs with a class of losses, including the hinge loss, all local minima are global minima with zero loss errors, using random matrix theory.

\*\*\*\*\*  
An experimental study of layer-level training speed and its impact on generalization

Simon Carbonnelle, Christophe De Vleeschouwer

How optimization influences the generalization ability of a DNN is still an active area of research. This work aims to unveil and study a factor of influence: the speed at which each layer trains. In our preliminary work, we develop a visualization technique and an optimization algorithm to monitor and control the layer rotation rate, a tentative measure of layer-level training speed, and show that it has a remarkably consistent and substantial impact on generalization. Our experiments further suggest that weight decay's and adaptive gradients methods' impact on both generalization performance and speed of convergence are solely due to layer rotation rate changes compared to vanilla SGD, offering a novel interpretation of these widely used techniques, and providing supplementary evidence that layer-level training speed indeed impacts generalization. Besides these fundamental findings, we also expect that on a practical level, the tools we introduce will reduce the meta-parameter tuning required to get the best generalization out of a deep network.

\*\*\*\*\*

MERCI: A NEW METRIC TO EVALUATE THE CORRELATION BETWEEN PREDICTIVE UNCERTAINTY AND TRUE ERROR

michel moukari,loïc simon,sylvaine picard,frédéric jurie

As deep learning applications are becoming more and more pervasive, the question of evaluating the reliability of a prediction becomes a central question in the machine learning community. This domain, known as predictive uncertainty, has come under the scrutiny of research groups developing Bayesian approaches to deep learning such as Monte Carlo Dropout. Unfortunately, for the time being, the real goal of predictive uncertainty has been swept under the rug. Indeed, Bayesian approaches are solely evaluated in terms of raw performance of the prediction, while the quality of the estimated uncertainty is not assessed. One contribution of this article is to draw attention on existing metrics developed in the forecasting community, designed to evaluate both the sharpness and the calibration of predictive uncertainty. Sharpness refers to the concentration of the predictive distributions and calibration to the consistency between the predicted uncertainty level and the actual errors. We further analyze the behavior of these metrics on regression problems when deep convolutional networks are involved and for several current predictive uncertainty approaches. A second contribution of this article is to propose an alternative metric that is more adapted to the evaluation of relative uncertainty assessment and directly applicable to regression with deep learning. This metric is evaluated and compared with existing ones on a toy dataset as well as on the problem of monocular depth estimation.

\*\*\*\*\*

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

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

ex.

\*\*\*\*\*

### Selective Self-Training for semi-supervised Learning

Jisoo Jeong, Seungeui Lee, Nojun Kwak

Semi-supervised learning (SSL) is a study that efficiently exploits a large amount of unlabeled data to improve performance in conditions of limited labeled data. Most of the conventional SSL methods assume that the classes of unlabeled data are included in the set of classes of labeled data. In addition, these methods do not sort out useless unlabeled samples and use all the unlabeled data for learning, which is not suitable for realistic situations. In this paper, we propose an SSL method called selective self-training (SST), which selectively decides whether to include each unlabeled sample in the training process. It is also designed to be applied to a more real situation where classes of unlabeled data are different from the ones of the labeled data. For the conventional SSL problems which deal with data where both the labeled and unlabeled samples share the same class categories, the proposed method not only performs comparable to other conventional SSL algorithms but also can be combined with other SSL algorithms. While the conventional methods cannot be applied to the new SSL problems where the separated data do not share the classes, our method does not show any performance degradation even if the classes of unlabeled data are different from those of the labeled data.

\*\*\*\*\*

### Feature Attribution As Feature Selection

Satoshi Hara, Koichi Ikeno, Tasuku Soma, Takanori Maehara

Feature attribution methods identify "relevant" features as an explanation of a complex machine learning model. Several feature attribution methods have been proposed; however, only a few studies have attempted to define the "relevance" of each feature mathematically. In this study, we formalize the feature attribution problem as a feature selection problem. In our proposed formalization, there arise two possible definitions of relevance. We name the feature attribution problems based on these two relevances as Exclusive Feature Selection (EFS) and Inclusive Feature Selection (IFS). We show that several existing feature attribution methods can be interpreted as approximation algorithms for EFS and IFS. Moreover, through exhaustive experiments, we show that IFS is better suited as the formalization for the feature attribution problem than EFS.

\*\*\*\*\*

### Gradient-based Training of Slow Feature Analysis by Differentiable Approximate Whitening

Merlin Schüler, Hlynur Dávid Hlynsson, Laurenz Wiskott

We propose Power Slow Feature Analysis, a gradient-based method to extract temporally slow features from a high-dimensional input stream that varies on a faster time-scale, as a variant of Slow Feature Analysis (SFA). While displaying performance comparable to hierarchical extensions to the SFA algorithm, such as Hierarchical Slow Feature Analysis, for a small number of output-features, our algorithm allows fully differentiable end-to-end training of arbitrary differentiable approximators (e.g., deep neural networks). We provide experimental evidence that PowerSFA is able to extract meaningful and informative low-dimensional features in the case of (a) synthetic low-dimensional data, (b) visual data, and also for (c) a general dataset for which symmetric non-temporal relations between points can be defined.

\*\*\*\*\*

### SEQUENCE MODELLING WITH AUTO-ADDRESSING AND RECURRENT MEMORY INTEGRATING NETWORKS

Zhangheng Li, Jia-Xing Zhong, Jingjia Huang, Tao Zhang, Thomas Li, Ge Li

Processing sequential data with long term dependencies and learn complex transitions are two major challenges in many deep learning applications. In this paper, we introduce a novel architecture, the Auto-addressing and Recurrent Memory Integrating Network (ARMIN) to address these issues. The ARMIN explicitly stores previous hidden states and recurrently integrate useful past states into current time-step by an efficient memory addressing mechanism. Compared to existing memor

y networks, the ARMIN is more light-weight and inference-time efficient. Our network can be trained on small slices of long sequential data, and thus, can boost its training speed. Experiments on various tasks demonstrate the efficiency of the ARMIN architecture. Codes and models will be available.

\*\*\*\*\*

SHE2: Stochastic Hamiltonian Exploration and Exploitation for Derivative-Free Optimization

Haoyi Xiong, Wenqing Hu, Zhanxing Zhu, Xinjian Li, Yunchao Zhang, Jun Huan

Derivative-free optimization (DFO) using trust region methods is frequently used for machine learning applications, such as (hyper-)parameter optimization without the derivatives of objective functions known. Inspired by the recent work in continuous-time minimizers, our work models the common trust region methods with the exploration-exploitation using a dynamical system coupling a pair of dynamical processes. While the first exploration process searches the minimum of the blackbox function through minimizing a time-evolving surrogation function, another exploitation process updates the surrogation function time-to-time using the points traversed by the exploration process. The efficiency of derivative-free optimization thus depends on ways the two processes couple. In this paper, we propose a novel dynamical system, namely Stochastic Hamiltonian Exploration and Exploitation, that surrogates the subregions of blackbox function using a time-evolving quadratic function, then explores and tracks the minimum of the quadratic functions using a fast-converging Hamiltonian system. The Algorithm is later provided as a discrete-time numerical approximation to the system. To further accelerate optimization, we present Parallel that parallelizes multiple Algorithm threads for concurrent exploration and exploitation. Experiment results based on a wide range of machine learning applications show that Parallel outperform a boarder range of derivative-free optimization algorithms with faster convergence speed under the same settings.

\*\*\*\*\*

Image Score: how to select useful samples

Simiao Zuo, Jialin Wu

There has long been debates on how we could interpret neural networks and understand the decisions our models make. Specifically, why deep neural networks tend to be error-prone when dealing with samples that output low softmax scores. We present an efficient approach to measure the confidence of decision-making steps by statistically investigating each unit's contribution to that decision. Instead of focusing on how the models react on datasets, we study the datasets themselves given a pre-trained model. Our approach is capable of assigning a score to each sample within a dataset that measures the frequency of occurrence of that sample's chain of activation. We demonstrate with experiments that our method could select useful samples to improve deep neural networks in a semi-supervised learning setting.

\*\*\*\*\*

Improved resistance of neural networks to adversarial images through generative pre-training

Joachim Wabnig

We train a feed forward neural network with increased robustness against adversarial attacks compared to conventional training approaches. This is achieved using a novel pre-trained building block based on a mean field description of a Boltzmann machine. On the MNIST dataset the method achieves strong adversarial resistance without data augmentation or adversarial training. We show that the increased adversarial resistance is correlated with the generative performance of the underlying Boltzmann machine.

\*\*\*\*\*

Sample Efficient Deep Neuroevolution in Low Dimensional Latent Space

Bin Zhou, Jiashi Feng

Current deep neuroevolution models are usually trained in a large parameter search space for complex learning tasks, e.g. playing video games, which needs billions of samples and thousands of search steps to obtain significant performance.

This raises a question of whether we can make use of sequential data generated during evolution, encode input samples, and evolve in low dimensional parameter space with latent state input in a fast and efficient manner. Here we give an affirmative answer: we train a VAE to encode input samples, then an RNN to model environment dynamics and handle temporal information, and last evolve our low dimensional policy network in latent space. We demonstrate that this approach is surprisingly efficient: our experiments on Atari games show that within 10M frames and 30 evolution steps of training, our algorithm could achieve competitive result compared with ES, A3C, and DQN which need billions of frames.

\*\*\*\*\*

#### GRAPH TRANSFORMATION POLICY NETWORK FOR CHEMICAL REACTION PREDICTION

Kien Do,Truyen Tran,Svetha Venkatesh

We address a fundamental problem in chemistry known as chemical reaction product prediction. Our main insight is that the input reactant and reagent molecules can be jointly represented as a graph, and the process of generating product molecules from reactant molecules can be formulated as a sequence of graph transformations. To this end, we propose Graph Transformation Policy Network (GTPN) - a novel generic method that combines the strengths of graph neural networks and reinforcement learning to learn the reactions directly from data with minimal chemical knowledge. Compared to previous methods, GTPN has some appealing properties such as: end-to-end learning, and making no assumption about the length or the order of graph transformations. In order to guide model search through the complex discrete space of sets of bond changes effectively, we extend the standard policy gradient loss by adding useful constraints. Evaluation results show that GTPN improves the top-1 accuracy over the current state-of-the-art method by about 3% on the large USPTO dataset. Our model's performances and prediction errors are also analyzed carefully in the paper.

\*\*\*\*\*

#### On the Convergence and Robustness of Batch Normalization

Yongqiang Cai,Qianxiao Li,Zuwei Shen

Despite its empirical success, the theoretical underpinnings of the stability, convergence and acceleration properties of batch normalization (BN) remain elusive. In this paper, we attack this problem from a modelling approach, where we perform thorough theoretical analysis on BN applied to simplified model: ordinary least squares (OLS). We discover that gradient descent on OLS with BN has interesting properties, including a scaling law, convergence for arbitrary learning rates for the weights, asymptotic acceleration effects, as well as insensitivity to choice of learning rates. We then demonstrate numerically that these findings are not specific to the OLS problem and hold qualitatively for more complex supervised learning problems. This points to a new direction towards uncovering the mathematical principles that underlies batch normalization.

\*\*\*\*\*

#### Learning Neuron Non-Linearities with Kernel-Based Deep Neural Networks

Giuseppe Marra,Dario Zanca,Alessandro Betti,Marco Gori

The effectiveness of deep neural architectures has been widely supported in terms of both experimental and foundational principles. There is also clear evidence that the activation function (e.g. the rectifier and the LSTM units) plays a crucial role in the complexity of learning. Based on this remark, this paper discusses an optimal selection of the neuron non-linearity in a functional framework that is inspired from classic regularization arguments. A representation theorem is given which indicates that the best activation function is a kernel expansion in the training set, that can be effectively approximated over an opportune set of points modeling 1-D clusters. The idea can be naturally extended to recurrent networks, where the expressiveness of kernel-based activation functions turns out to be a crucial ingredient to capture long-term dependencies. We give experimental evidence of this property by a set of challenging experiments, where we compare the results with neural architectures based on state of the art LSTM cells.

\*\*\*\*\*

#### Human Action Recognition Based on Spatial-Temporal Attention

Wensong Chan,Zhiqiang Tian,Xuguang Lan

Many state-of-the-art methods of recognizing human action are based on attention mechanism, which shows the importance of attention mechanism in action recognition. With the rapid development of neural networks, human action recognition has been achieved great improvement by using convolutional neural networks (CNN) or recurrent neural networks (RNN). In this paper, we propose a model based on spatial-temporal attention weighted LSTM. This model pays attention to the key part in each video frame, and also focuses on the important frames in each video sequence, thus the most important theme for our model is how to find out the key point spatially and the key frames temporally. We show a feasible architecture which can solve those two problems effectively and achieve a satisfactory result. Our model is trained and tested on three datasets including UCF-11, UCF-101, and HMDB51. Those results demonstrate a high performance of our model in human action recognition.

\*\*\*\*\*

FAVAE: SEQUENCE DISENTANGLEMENT USING INFORMATION BOTTLENECK PRINCIPLE

Masanori Yamada, Kim Heecheol, Kosuke Miyoshi, Hiroshi Yamakawa

A state-of-the-art generative model, a "factorized action variational autoencoder (FAVAE)," is presented for learning disentangled and interpretable representations from sequential data via the information bottleneck without supervision. The purpose of disentangled representation learning is to obtain interpretable and transferable representations from data. We focused on the disentangled representation of sequential data because there is a wide range of potential applications if disentanglement representation is extended to sequential data such as video, speech, and stock price data. Sequential data is characterized by dynamic factors and static factors: dynamic factors are time-dependent, and static factors are independent of time. Previous works succeed in disentangling static factors and dynamic factors by explicitly modeling the priors of latent variables to distinguish between static and dynamic factors. However, this model can not disentangle representations between dynamic factors, such as disentangling "picking" and "throwing" in robotic tasks. In this paper, we propose new model that can disentangle multiple dynamic factors. Since our method does not require modeling priors, it is capable of disentangling "between" dynamic factors. In experiments, we show that FAVAE can extract the disentangled dynamic factors.

\*\*\*\*\*

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

Andrew K. Lampinen, Surya Ganguli

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 provably 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.

\*\*\*\*\*

#### DATNet: Dual Adversarial Transfer for Low-resource Named Entity Recognition

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Rick Siow Mong Goh, Kenneth Kwok

We propose a new architecture termed Dual Adversarial Transfer Network (DATNet) for addressing low-resource Named Entity Recognition (NER). Specifically, two variants of DATNet, i.e., DATNet-F and DATNet-P, are proposed to explore effective feature fusion between high and low resource. To address the noisy and imbalanced training data, we propose a novel Generalized Resource-Adversarial Discriminator (GRAD). Additionally, adversarial training is adopted to boost model generalization. We examine the effects of different components in DATNet across domains and languages and show that significant improvement can be obtained especially for low-resource data. Without augmenting any additional hand-crafted features, we achieve new state-of-the-art performances on CoNLL and Twitter NER---88.16% F1 for Spanish, 53.43% F1 for WNUT-2016, and 42.83% F1 for WNUT-2017.

\*\*\*\*\*

#### HIGHLY EFFICIENT 8-BIT LOW PRECISION INFERENCE OF CONVOLUTIONAL NEURAL NETWORKS

Haihao Shen, Jiong Gong, Xiaoli Liu, Guoming Zhang, Ge Jin, and Eric Lin

High throughput and low latency inference of deep neural networks are critical for the deployment of deep learning applications. This paper presents a general technique toward 8-bit low precision inference of convolutional neural networks, including 1) channel-wise scale factors of weights, especially for depthwise convolution, 2) Winograd convolution, and 3) topology-wise 8-bit support. We experiment the techniques on top of a widely-used deep learning framework. The 8-bit optimized model is automatically generated with a calibration process from FP32 model without the need of fine-tuning or retraining. We perform a systematical and comprehensive study on 18 widely-used convolutional neural networks and demonstrate the effectiveness of 8-bit low precision inference across a wide range of applications and use cases, including image classification, object detection, image segmentation, and super resolution. We show that the inference throughput and latency are improved by 1.6X and 1.5X respectively with minimal within 0.6% to no loss in accuracy from FP32 baseline. We believe the methodology can provide the guidance and reference design of 8-bit low precision inference for other frameworks. All the code and models will be publicly available soon.

\*\*\*\*\*

#### HC-Net: Memory-based Incremental Dual-Network System for Continual learning

Jangho Kim, Jeessoo Kim, Nojun Kwak

Training a neural network for a classification task typically assumes that the data to train are given from the beginning.

However, in the real world, additional data accumulate gradually and the model requires additional training without accessing the old training data. This usually leads to the catastrophic forgetting problem which is inevitable for the traditional training methodology of neural networks.

In this paper, we propose a memory-based continual learning method that is able to learn additional tasks while retaining the performance of previously learned tasks.

Composed of two complementary networks, the Hippocampus-Net (H-Net) and the Cortex-Net (C-Net), our model estimates the index of the corresponding task for an input sample and utilizes a particular portion of itself with the estimated index.

The C-Net guarantees no degradation in the performance of the previously learned tasks and the H-Net shows high confidence in finding the origin of an input sample.

\*\*\*\*\*

#### Are adversarial examples inevitable?

Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, Tom Goldstein

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



ntifies 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.

\*\*\*\*\*

SOSELETO: A Unified Approach to Transfer Learning and Training with Noisy Labels  
Or Litany, Daniel Freedman

We present SOSELETO (Source SELECTION for Target Optimization), a new method for exploiting a source dataset to solve a classification problem on a target dataset. SOSELETO is based on the following simple intuition: some source examples are more informative than others for the target problem. To capture this intuition, source samples are each given weights; these weights are solved for jointly with the source and target classification problems via a bilevel optimization scheme. The target therefore gets to choose the source samples which are most informative for its own classification task. Furthermore, the bilevel nature of the optimization acts as a kind of regularization on the target, mitigating overfitting. SOSELETO may be applied to both classic transfer learning, as well as the problem of training on datasets with noisy labels; we show state of the art results on both of these problems.

\*\*\*\*\*

Local Critic Training of Deep Neural Networks

Hojung Lee, Jong-Seok Lee

This paper proposes a novel approach to train deep neural networks by unlocking the layer-wise dependency of backpropagation training. The approach employs additional modules called local critic networks besides the main network model to be trained, which are used to obtain error gradients without complete feedforward and backward propagation processes. We propose a cascaded learning strategy for these local networks. In addition, the approach is also useful from multi-model perspectives, including structural optimization of neural networks, computationally efficient progressive inference, and ensemble classification for performance improvement. Experimental results show the effectiveness of the proposed approach and suggest guidelines for determining appropriate algorithm parameters.

\*\*\*\*\*

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

Mohit Sharma, Arjun Sharma, Nicholas Rhinehart, Kris M. Kitani

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<sup>3</sup>RL: Mind-aware Multi-agent Management Reinforcement Learning

Tianmin Shu, Yuandong Tian

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<sup>3</sup>RL), 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.

\*\*\*\*\*

CoT: Cooperative Training for Generative Modeling of Discrete Data

Sidi Lu, Lantao Yu, Siyuan Feng, Yaoming Zhu, Weinan Zhang, Yong Yu

We propose Cooperative Training (CoT) for training generative models that measure a tractable density for discrete data. CoT coordinately trains a generator  $G$  and an auxiliary predictive mediator  $M$ . The training target of  $M$  is to estimate a mixture density of the learned distribution  $G$  and the target distribution  $P$ , and that of  $G$  is to minimize the Jensen-Shannon divergence estimated through  $M$ . CoT achieves independent success without the necessity of pre-training via Maximum Likelihood Estimation or involving high-variance algorithms like REINFORCE. This low-variance algorithm is theoretically proved to be superior for both sample generation and likelihood prediction. We also theoretically and empirically show the superiority of CoT over most previous algorithms in terms of generative quality and diversity, predictive generalization ability and computational cost.

\*\*\*\*\*

An Exhaustive Analysis of Lazy vs. Eager Learning Methods for Real-Estate Property Investment

Setareh Rafatirad, Maryam Heidari

Accurate rent prediction in real estate investment can help in generating capital gains and guaranty a financial success. In this paper, we carry out a comprehensive analysis and study of eleven machine learning algorithms for rent prediction, including Linear Regression, Multilayer Perceptron, Random Forest, KNN, ML-KNN, Locally Weighted Learning, SMO, SVM, J48, lazy Decision Tree (i.e., lazy DT), and KStar algorithms.

Our contribution in this paper is twofold: (1) We present a comprehensive analysis of internal and external attributes of a real-estate housing dataset and their correlation with rental prices. (2) We use rental prediction as a platform to study and compare the performance of eager vs. lazy machine learning methods using myriad of ML algorithms.

We train our rent prediction models using a Zillow data set of 4K real estate properties in Virginia State of the US, including three house types of single-family, townhouse, and condo. Each data instance in the dataset has 21 internal attributes (e.g., area space, price, number of bed/bath, rent, school rating, so forth). In addition to Zillow data, external attributes like walk/transit score, and crime rate are collected from online data sources. A subset of the collected features - determined by the PCA technique - are selected to tune the parameters of the prediction models. We employ a hierarchical clustering approach to cluster the data based on two factors of house type, and average rent estimate of zip codes. We evaluate and compare the efficacy of the tuned prediction models based on two metrics of R-squared and Mean Absolute Error, applied on unseen data. Based on our study, lazy models like KStar lead to higher accuracy and lower prediction error compared to eager methods like J48 and LR. However, it is not necessarily found to be an overarching conclusion drawn from the comparison between all the lazy and eager methods in this work.

\*\*\*\*\*

Differentiable Learning-to-Normalize via Switchable Normalization

Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, Jingyu Li

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.

\*\*\*\*\*

Learning a Neural-network-based Representation for Open Set Recognition

Mehadi Hassen, Philip K. Chan

In this paper, we present a neural network based representation for addressing the open set recognition problem. In this representation instances from the same class are close to each other while instances from different classes are further apart, resulting in statistically significant improvement when compared to other approaches on three datasets from two different domains.

\*\*\*\*\*

Learning to Reinforcement Learn by Imitation

Rosen Kralev, Russell Mendonca, Alvin Zhang, Tianhe Yu, Abhishek Gupta, Pieter Abbeel, Sergey Levine, Chelsea Finn

Meta-reinforcement learning aims to learn fast reinforcement learning (RL) procedures that can be applied to new tasks or environments. While learning fast RL procedures holds promise for allowing agents to autonomously learn a diverse range of skills, existing methods for learning efficient RL are impractical for real world settings, as they rely on slow reinforcement learning algorithms for meta-training, even when the learned procedures are fast. In this paper, we propose to learn a fast reinforcement learning procedure through supervised imitation of an expert, such that, after meta-learning, an agent can quickly learn new tasks through trial-and-error. Through our proposed method, we show that it is possible to learn fast RL using demonstrations, rather than relying on slow RL, where expert agents can be trained quickly by using privileged information or off-policy RL methods. Our experimental evaluation on a number of complex simulated robotic domains demonstrates that our method can effectively learn to learn from sparse rewards and is significantly more efficient than prior meta reinforcement learning algorithms.

\*\*\*\*\*

Learning Latent Semantic Representation from Pre-defined Generative Model

Jin-Young Kim, Sung-Bae Cho

Learning representations of data is an important issue in machine learning. Though GAN has led to significant improvements in the data representations, it still has several problems such as unstable training, hidden manifold of data, and huge computational overhead. GAN tends to produce the data simply without any information about the manifold of the data, which hinders from controlling desired features to generate. Moreover, most of GAN's have a large size of manifold, resulting in poor scalability. In this paper, we propose a novel GAN to control the latent semantic representation, called LSC-GAN, which allows us to produce desired data to generate and learns a representation of the data efficiently. Unlike the conventional GAN models with hidden distribution of latent space, we define the distributions explicitly in advance that are trained to generate the data based on the corresponding features by inputting the latent variables that follow the distribution. As the larger scale of latent space caused by deploying various distributions in one latent space makes training unstable while maintaining the

e dimension of latent space, we need to separate the process of defining the distributions explicitly and operation of generation. We prove that a VAE is proper for the former and modify a loss function of VAE to map the data into the pre-defined latent space so as to locate the reconstructed data as close to the input data according to its characteristics. Moreover, we add the KL divergence to the loss function of LSC-GAN to include this process. The decoder of VAE, which generates the data with the corresponding features from the pre-defined latent space, is used as the generator of the LSC-GAN. Several experiments on the CelebA dataset are conducted to verify the usefulness of the proposed method to generate desired data stably and efficiently, achieving a high compression ratio that can hold about 24 pixels of information in each dimension of latent space. Besides, our model learns the reverse of features such as not laughing (rather frowning) only with data of ordinary and smiling facial expression.

\*\*\*\*\*

q-Neurons: Neuron Activations based on Stochastic Jackson's Derivative Operators  
Frank Nielsen, Ke Sun

We propose a new generic type of stochastic neurons, called  $q$ -neurons, that considers activation functions based on Jackson's  $q$ -derivatives, with stochastic parameters  $q$ . Our generalization of neural network architectures with  $q$ -neurons is shown to be both scalable and very easy to implement. We demonstrate experimentally consistently improved performances over state-of-the-art standard activation functions, both on training and testing loss functions.

\*\*\*\*\*

Backplay: 'Man muss immer umkehren'

Cinjon Resnick, Roberta Raileanu, Sanyam Kapoor, Alexander Peysakhovich, Kyunghyun Cho, Joan Bruna

Model-free reinforcement learning (RL) requires a large number of trials to learn a good policy, especially in environments with sparse rewards. We explore a method to improve the sample efficiency when we have access to demonstrations. Our approach, Backplay, uses a single demonstration to construct a curriculum for a given task. Rather than starting each training episode in the environment's fixed initial state, we start the agent near the end of the demonstration and move the starting point backwards during the course of training until we reach the initial state. Our contributions are that we analytically characterize the types of environments where Backplay can improve training speed, demonstrate the effectiveness of Backplay both in large grid worlds and a complex four player zero-sum game (Pommerman), and show that Backplay compares favorably to other competitive methods known to improve sample efficiency. This includes reward shaping, behavioral cloning, and reverse curriculum generation.

\*\*\*\*\*

Using Ontologies To Improve Performance In Massively Multi-label Prediction

Ethan Steinberg, Peter J. Liu

Massively multi-label prediction/classification problems arise in environments like health-care or biology where it is useful to make very precise predictions. One challenge with massively multi-label problems is that there is often a long-tailed frequency distribution for the labels, resulting in few positive examples for the rare labels. We propose a solution to this problem by modifying the output layer of a neural network to create a Bayesian network of sigmoids which takes advantage of ontology relationships between the labels to help share information between the rare and the more common labels. We apply this method to the two massively multi-label tasks of disease prediction (ICD-9 codes) and protein function prediction (Gene Ontology terms) and obtain significant improvements in per-label AUROC and average precision.

\*\*\*\*\*

Boosting Trust Region Policy Optimization by Normalizing flows Policy

Yunhao Tang, Shipra Agrawal

We propose to improve trust region policy search with normalizing flows policy. We illustrate that when the trust region is constructed by KL divergence constraint, normalizing flows policy can generate samples far from the 'center' of the

previous policy iterate, which potentially enables better exploration and helps avoid bad local optima. We show that normalizing flows policy significantly improves upon factorized Gaussian policy baseline, with both TRPO and ACKTR, especially on tasks with complex dynamics such as Humanoid.

\*\*\*\*\*

Like What You Like: Knowledge Distill via Neuron Selectivity Transfer

Zehao Huang, Naiyan Wang

Despite deep neural networks have demonstrated extraordinary power in various applications, their superior performances are at expense of high storage and computational costs. Consequently, the acceleration and compression of neural networks have attracted much attention recently. Knowledge Transfer (KT), which aims at training a smaller student network by transferring knowledge from a larger teacher model, is one of the popular solutions. In this paper, we propose a novel knowledge transfer method by treating it as a distribution matching problem. Particularly, we match the distributions of neuron selectivity patterns between teacher and student networks. To achieve this goal, we devise a new KT loss function by minimizing the Maximum Mean Discrepancy (MMD) metric between these distributions. Combined with the original loss function, our method can significantly improve the performance of student networks. We validate the effectiveness of our method across several datasets, and further combine it with other KT methods to explore the best possible results. Last but not least, we fine-tune the model to other tasks such as object detection. The results are also encouraging, which confirm the transferability of the learned features.

\*\*\*\*\*

Massively Parallel Hyperparameter Tuning

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Ben Recht, Ameet Talwalkar

Modern learning models are characterized by large hyperparameter spaces. In order to adequately explore these large spaces, we must evaluate a large number of configurations, typically orders of magnitude more configurations than available parallel workers. Given the growing costs of model training, we would ideally like to perform this search in roughly the same wall-clock time needed to train a single model. In this work, we tackle this challenge by introducing ASHA, a simple and robust hyperparameter tuning algorithm with solid theoretical underpinnings that exploits parallelism and aggressive early-stopping. Our extensive empirical results show that ASHA outperforms state-of-the-art hyperparameter tuning methods; scales linearly with the number of workers in distributed settings; converges to a high quality configuration in half the time taken by Vizier, Google's internal hyperparameter tuning service) in an experiment with 500 workers; and beats the published result for a near state-of-the-art LSTM architecture in under  $2\times$  the time to train a single model.

\*\*\*\*\*

Unsupervised Meta-Learning for Reinforcement Learning

Abhishek Gupta, Benjamin Eysenbach, Chelsea Finn, Sergey Levine

Meta-learning is a powerful tool that learns how to quickly adapt a model to new tasks. In the context of reinforcement learning, meta-learning algorithms can acquire reinforcement learning procedures to solve new problems more efficiently by meta-learning prior tasks. The performance of meta-learning algorithms critically depends on the tasks available for meta-training: in the same way that supervised learning algorithms generalize best to test points drawn from the same distribution as the training points, meta-learning methods generalize best to tasks from the same distribution as the meta-training tasks. In effect, meta-reinforcement learning offloads the design burden from algorithm design to task design.

If we can automate the process of task design as well, we can devise a meta-learning algorithm that is truly automated. In this work, we take a step in this direction, proposing a family of unsupervised meta-learning algorithms for reinforcement learning. We describe a general recipe for unsupervised meta-reinforcement learning, and describe an effective instantiation of this approach based on a recently proposed unsupervised exploration technique and model-agnostic meta-learning. We also discuss practical and conceptual considerations for developing un

supervised meta-learning methods. Our experimental results demonstrate that supervised meta-reinforcement learning effectively acquires accelerated reinforcement learning procedures without the need for manual task design, significantly exceeds the performance of learning from scratch, and even matches performance of meta-learning methods that use hand-specified task distributions.

\*\*\*\*\*

#### Stochastic Learning of Additive Second-Order Penalties with Applications to Fairness

Heinrich Jiang, Yifan Wu, Ofir Nachum

Many notions of fairness may be expressed as linear constraints, and the resulting constrained objective is often optimized by transforming the problem into its Lagrangian dual with additive linear penalties. In non-convex settings, the resulting problem may be difficult to solve as the Lagrangian is not guaranteed to have a deterministic saddle-point equilibrium. In this paper, we propose to modify the linear penalties to second-order ones, and we argue that this results in a more practical training procedure in non-convex, large-data settings. For one, the use of second-order penalties allows training the penalized objective with a fixed value of the penalty coefficient, thus avoiding the instability and potential lack of convergence associated with two-player min-max games. Secondly, we derive a method for efficiently computing the gradients associated with the second-order penalties in stochastic mini-batch settings. Our resulting algorithm performs well empirically, learning an appropriately fair classifier on a number of standard benchmarks.

\*\*\*\*\*

#### Supervised Policy Update for Deep Reinforcement Learning

Quan Vuong, Yiming Zhang, Keith W. Ross

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

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.

\*\*\*\*\*

Dual Skew Divergence Loss for Neural Machine Translation

Yingting Wu,Hai Zhao,Rui Wang

For neural sequence model training, maximum likelihood (ML) has been commonly adopted to optimize model parameters with respect to the corresponding objective. However, in the case of sequence prediction tasks like neural machine translation (NMT), training with the ML-based cross entropy loss would often lead to models that overgeneralize and plunge into local optima. In this paper, we propose an extended loss function called dual skew divergence (DSD), which aims to give a better tradeoff between generalization ability and error avoidance during NMT training. Our empirical study indicates that switching to DSD loss after the convergence of ML training helps the model skip the local optimum and stimulates a stable performance improvement. The evaluations on WMT 2014 English-German and English-French translation tasks demonstrate that the proposed loss indeed helps bring about better translation performance than several baselines.

\*\*\*\*\*

Unified recurrent network for many feature types

Alexander Stec,Diego Klabjan,Jean Utke

There are time series that are amenable to recurrent neural network (RNN) solutions when treated as sequences, but some series, e.g. asynchronous time series, provide a richer variation of feature types than current RNN cells take into account. In order to address such situations, we introduce a unified RNN that handles five different feature types, each in a different manner. Our RNN framework separates sequential features into two groups dependent on their frequency, which we call sparse and dense features, and which affect cell updates differently. Further, we also incorporate time features at the sequential level that relate to the time between specified events in the sequence and are used to modify the cell's memory state. We also include two types of static (whole sequence level) features, one related to time and one not, which are combined with the encoder output. The experiments show that the proposed modeling framework does increase performance compared to standard cells.

\*\*\*\*\*

Gradient Descent Happens in a Tiny Subspace

Guy Gur-Ari,Daniel A. Roberts,Ethan Dyer

We show that in a variety of large-scale deep learning scenarios the gradient dynamically converges to a very small subspace after a short period of training. The subspace is spanned by a few top eigenvectors of the Hessian (equal to the number of classes in the dataset), and is mostly preserved over long periods of training. A simple argument then suggests that gradient descent may happen mostly in this subspace. We give an example of this effect in a solvable model of classification, and we comment on possible implications for optimization and learning.

\*\*\*\*\*

Improved Learning of One-hidden-layer Convolutional Neural Networks with Overlaps

Simon S. Du,Surbhi Goel

We propose a new algorithm to learn a one-hidden-layer convolutional neural network where both the convolutional weights and the outputs weights are parameters to be learned. Our algorithm works for a general class of (potentially overlapping) patches, including commonly used structures for computer vision tasks. Our algorithm draws ideas from (1) isotonic regression for learning neural networks and (2) landscape analysis of non-convex matrix factorization problems. We believe these findings may inspire further development in designing provable algorithms for learning neural networks and other complex models. While our focus is theoretical, we also present experiments that illustrate our theoretical findings.

\*\*\*\*\*

#### LEARNING FACTORIZED REPRESENTATIONS FOR OPEN-SET DOMAIN ADAPTATION

Mahsa Baktashmotlagh, Masoud Faraki, Tom Drummond, Mathieu Salzmann

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.

\*\*\*\*\*

#### Spread Divergences

David Barber, Mingtian Zhang, Raza Habib, Thomas Bird

For distributions  $p$  and  $q$  with different support, the divergence  $\text{div}\{p\}\{q\}$  generally will not exist. We define a spread divergence  $\text{sdiv}\{p\}\{q\}$  on modified  $p$  and  $q$  and describe sufficient conditions for the existence of such a divergence. We give examples of using a spread divergence to train implicit generative models, including linear models (Principal Components Analysis and Independent Components Analysis) and non-linear models (Deep Generative Networks).

\*\*\*\*\*

#### Adversarial Attacks for Optical Flow-Based Action Recognition Classifiers

Nathan Inkawhich, Matthew Inkawhich, Hai Li, Yiran Chen

The success of deep learning research has catapulted deep models into production systems that our society is becoming increasingly dependent on, especially in the

image and video domains. However, recent work has shown that these largely uninterpretable models exhibit glaring security vulnerabilities in the presence of

an adversary. In this work, we develop a powerful untargeted adversarial attack for action recognition systems in both white-box and black-box settings. Action recognition models differ from image-classification models in that their inputs contain a temporal dimension, which we explicitly target in the attack. Drawing inspiration from image classifier attacks, we create new attacks which achieve state-of-the-art success rates on a two-stream classifier trained on the UCF-101 dataset. We find that our attacks can significantly degrade a model's performance

with sparsely and imperceptibly perturbed examples. We also demonstrate the transferability of our attacks to black-box action recognition systems.

\*\*\*\*\*

#### TequilaGAN: How To Easily Identify GAN Samples

Rafael Valle, Wilson Cai, Anish P. Doshi

In this paper we show strategies to easily identify fake samples generated with the Generative Adversarial Network framework. One strategy is based on the stati



stical analysis and comparison of raw pixel values and features extracted from them. The other strategy learns formal specifications from the real data and shows that fake samples violate the specifications of the real data. We show that fake samples produced with GANs have a universal signature that can be used to identify fake samples. We provide results on MNIST, CIFAR10, music and speech data.  
\*\*\*\*\*

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

Florian Tramer, Dan Boneh

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.

\*\*\*\*\*

MANIFOLDNET: A DEEP NEURAL NETWORK FOR MANIFOLD-VALUED DATA

Rudrasis Chakraborty, Jose Bouza, Jonathan Manton, Baba C. Vemuri

Developing deep neural networks (DNNs) for manifold-valued data sets has gained much interest of late in the deep learning research community. Examples of manifold-valued data include data from omnidirectional cameras on automobiles, drones etc., diffusion magnetic resonance imaging, elastography and others. In this paper, we present a novel theoretical framework for DNNs to cope with manifold-valued data inputs. In doing this generalization, we draw parallels to the widely popular convolutional neural networks (CNNs). We call our network the ManifoldNet.

As in vector spaces where convolutions are equivalent to computing the weighted mean of functions, an analogous definition for manifold-valued data can be constructed involving the computation of the weighted Fréchet Mean (wFM). To this end, we present a provably convergent recursive computation of the wFM of the given data, where the weights make up the convolution mask, to be learned. Further, we prove that the proposed wFM layer achieves a contraction mapping and hence the ManifoldNet does not need the additional non-linear ReLU unit used in standard CNNs. Operations such as pooling in traditional CNN are no longer necessary in this setting since wFM is already a pooling type operation. Analogous to the equivariance of convolution in Euclidean space to translations, we prove that the wFM is equivariant to the action of the group of isometries admitted by the Riemannian manifold on which the data reside. This equivariance property facilitates weight sharing within the network. We present experiments, using the ManifoldNet framework, to achieve video classification and image reconstruction using an auto-encoder+decoder setting. Experimental results demonstrate the efficacy of ManifoldNet in the context of classification and reconstruction accuracy.

\*\*\*\*\*

Effective and Efficient Batch Normalization Using Few Uncorrelated Data for Statistics' Estimation

Zhaodong Chen, Lei Deng, Guoqi Li, Jiawei Sun, Xing Hu, Ling Liang, Yufei Ding, Yuan Xie

Deep Neural Networks (DNNs) thrive in recent years in which Batch Normalization (BN) plays an indispensable role. However, it has been observed that BN is costly due to the reduction operations. In this paper, we propose alleviating the BN's cost by using only a small fraction of data for mean & variance estimation at each iteration. The key challenge to reach this goal is how to achieve a satisfactory balance between normalization effectiveness and execution efficiency. We identify that the effectiveness expects less data correlation while the efficiency expects regular execution pattern. To this end, we propose two categories of a approach: sampling or creating few uncorrelated data for statistics' estimation with certain strategy constraints. The former includes "Batch Sampling (BS)" that randomly selects few samples from each batch and "Feature Sampling (FS)" that randomly selects a small patch from each feature map of all samples, and the latter is "Virtual Dataset Normalization (VDN)" that generates few synthetic random samples. Accordingly, multi-way strategies are designed to reduce the data correlation for accurate estimation and optimize the execution pattern for running acceleration in the meantime. All the proposed methods are comprehensively evaluated on various DNN models, where an overall training speedup by up to 21.7% on modern GPUs can be practically achieved without the support of any specialized libraries, and the loss of model accuracy and convergence rate are negligible. Furthermore, our methods demonstrate powerful performance when solving the well-known "micro-batch normalization" problem in the case of tiny batch size.

\*\*\*\*\*

Learning to Understand Goal Specifications by Modelling Reward

Dzmitry Bahdanau, Felix Hill, Jan Leike, Edward Hughes, Arian Hosseini, Pushmeet Kohli, Edward Grefenstette

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.

\*\*\*\*\*

Reducing Overconfident Errors outside the Known Distribution

Zhizhong Li, Derek Hoiem

Intuitively, unfamiliarity should lead to lack of confidence. In reality, current algorithms often make highly confident yet wrong predictions when faced with unexpected test samples from an unknown distribution different from training. Unlike domain adaptation methods, we cannot gather an "unexpected dataset" prior to test, and unlike novelty detection methods, a best-effort original task prediction is still expected. We compare a number of methods from related fields such as calibration and epistemic uncertainty modeling, as well as two proposed methods that reduce overconfident errors of samples from an unknown novel distribution without drastically increasing evaluation time: (1) G-distillation, training an ensemble of classifiers and then distill into a single model using both labeled and unlabeled examples, or (2) NCR, reducing prediction confidence based on its novelty detection score. Experimentally, we investigate the overconfidence problem and evaluate our solution by creating "familiar" and "novel" test splits, where "familiar" are identically distributed with training and "novel" are not. We discover that calibrating using temperature scaling on familiar data is the best

t single-model method for improving novel confidence, followed by our proposed methods. In addition, some methods' NLL performance are roughly equivalent to a regularly trained model with certain degree of smoothing. Calibrating can also reduce confident errors, for example, in gender recognition by 95% on demographic groups different from the training data.

\*\*\*\*\*

#### Dynamic Sparse Graph for Efficient Deep Learning

Liu Liu, Lei Deng, Xing Hu, Maohua Zhu, Guoqi Li, Yufei Ding, Yuan Xie

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 dimension reduction 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

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.

\*\*\*\*\*

#### Adaptive Pruning of Neural Language Models for Mobile Devices

Raphael Tang, Jimmy Lin

Neural language models (NLMs) exist in an accuracy-efficiency tradeoff space where better perplexity typically comes at the cost of greater computation complexity. In a software keyboard application on mobile devices, this translates into higher power consumption and shorter battery life. This paper represents the first attempt, to our knowledge, in exploring accuracy-efficiency tradeoffs for NLMs. Building on quasi-recurrent neural networks (QRNNs), we apply pruning techniques to provide a "knob" to select different operating points. In addition, we propose a simple technique to recover some perplexity using a negligible amount of memory. Our empirical evaluations consider both perplexity as well as energy consumption on a Raspberry Pi, where we demonstrate which methods provide the best perplexity-power consumption operating point. At one operating point, one of the techniques is able to provide energy savings of 40% over the state of the art with only a 17% relative increase in perplexity.

\*\*\*\*\*

#### 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

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

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

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

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, regulariz

ing 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.

\*\*\*\*\*

#### GENERALIZED ADAPTIVE MOMENT ESTIMATION

Guoqiang Zhang, Kenta Niwa, W. Bastiaan Kleijn

Adaptive gradient methods have experienced great success in training deep neural networks (DNNs). The basic idea of the methods is to track and properly make use of the first and/or second moments of the gradient for model-parameter updates over iterations for the purpose of removing the need for manual interference. In this work, we propose a new adaptive gradient method, referred to as generalized adaptive moment estimation (Game). From a high level perspective, the new method introduces two more parameters w.r.t. AMSGrad (S. J. Reddi & Kumar (2018)) and one more parameter w.r.t. PAdam (Chen & Gu (2018)) to enlarge the parameter-selection space for performance enhancement while reducing the memory cost per iteration compared to AMSGrad and PAdam. The saved memory space amounts to the number of model parameters, which is significant for large-scale DNNs. Our motivation for introducing additional parameters in Game is to provide algorithmic flexibility to facilitate a reduction of the performance gap between training and validation datasets when training a DNN. Convergence analysis is provided for applying Game to solve both convex optimization and smooth nonconvex optimization. Empirical studies for training four convolutional neural networks over MNIST and CIFAR10 show that under proper parameter selection, Game produces promising validation performance as compared to AMSGrad and PAdam.

\*\*\*\*\*

#### DARTS: Differentiable Architecture Search

Hanxiao Liu, Karen Simonyan, Yiming Yang

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.

\*\*\*\*\*

#### Improved Gradient Estimators for Stochastic Discrete Variables

Evgeny Andriyash, Arash Vahdat, Bill Macready

In many applications we seek to optimize an expectation with respect to a distribution over discrete variables. Estimating gradients of such objectives with respect to the distribution parameters is a challenging problem. We analyze existing solutions including finite-difference (FD) estimators and continuous relaxation (CR) estimators in terms of bias and variance. We show that the commonly used Gumbel-Softmax estimator is biased and propose a simple method to reduce it. We also derive a simpler piece-wise linear continuous relaxation that also possesses reduced bias. We demonstrate empirically that reduced bias leads to a better performance in variational inference and on binary optimization tasks.

\*\*\*\*\*

#### Hybrid Policies Using Inverse Rewards for Reinforcement Learning

Yao Shi, Tian Xia, Guanjun Zhao, Xin Gao

This paper puts forward a broad-spectrum improvement for reinforcement learning algorithms, which combines the policies using original rewards and inverse (negative) rewards. The policies using inverse rewards are competitive with the original policies, and help the original policies correct their mis-actions. We have proved the convergence of the inverse policies. The experiments for some games i

n OpenAI gym show that the hybrid policies based on deep Q-learning, double Q-learning, and on-policy actor-critic obtain the rewards up to 63.8%, 97.8%, and 54.7% more than the original algorithms. The improved policies are more stable than the original policies as well.

\*\*\*\*\*

Generative Adversarial Network Training is a Continual Learning Problem

Kevin J Liang, Chunyuan Li, Guoyin Wang, Lawrence Carin

Generative Adversarial Networks (GANs) have proven to be a powerful framework for learning to draw samples from complex distributions. However, GANs are also notoriously difficult to train, with mode collapse and oscillations a common problem. We hypothesize that this is at least in part due to the evolution of the generator distribution and the catastrophic forgetting tendency of neural networks, which leads to the discriminator losing the ability to remember synthesized samples from previous instantiations of the generator. Recognizing this, our contributions are twofold. First, we show that GAN training makes for a more interesting and realistic benchmark for continual learning methods evaluation than some of the more canonical datasets. Second, we propose leveraging continual learning techniques to augment the discriminator, preserving its ability to recognize previous generator samples. We show that the resulting methods add only a light amount of computation, involve minimal changes to the model, and result in better overall performance on the examined image and text generation tasks.

\*\*\*\*\*

Identifying Generalization Properties in Neural Networks

Huan Wang, Nitish Shirish Keskar, Caiming Xiong, Richard Socher

While it has not yet been proven, empirical evidence suggests that model generalization is related to local properties of the optima which can be described via the Hessian. We connect model generalization with the local property of a solution under the PAC-Bayes paradigm. In particular, we prove that model generalization ability is related to the Hessian, the higher-order "smoothness" terms characterized by the Lipschitz constant of the Hessian, and the scales of the parameters. Guided by the proof, we propose a metric to score the generalization capability of the model, as well as an algorithm that optimizes the perturbed model accordingly.

\*\*\*\*\*

Adversarial Reprogramming of Neural Networks

Gamaleldin F. Elsayed, Ian Goodfellow, Jascha Sohl-Dickstein

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

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 sampli

ng 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

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.

\*\*\*\*\*

Unsupervised Word Discovery with Segmental Neural Language Models

Kazuya Kawakami, Chris Dyer, Phil Blunsom

We propose a segmental neural language model that combines the representational power of neural networks and the structure learning mechanism of Bayesian nonparametrics, and show that it learns to discover semantically meaningful units (e.g., morphemes and words) from unsegmented character sequences. The model generates text as a sequence of segments, where each segment is generated either character-by-character from a sequence model or as a single draw from a lexical memory that stores multi-character units. Its parameters are fit to maximize the marginal likelihood of the training data, summing over all segmentations of the input, and its hyperparameters are likewise set to optimize held-out marginal likelihood.

To prevent the model from overusing the lexical memory, which leads to poor generalization and bad segmentation, we introduce a differentiable regularizer that penalizes based on the expected length of each segment. To our knowledge, this is the first demonstration of neural networks that have predictive distributions better than LSTM language models and also infer a segmentation into word-like units that are competitive with the best existing word discovery models.

\*\*\*\*\*

Principled Deep Neural Network Training through Linear Programming

Daniel Bienstock, Gonzalo Muñoz, Sebastian Pokutta

Deep Learning has received significant attention due to its impressive performance in many state-of-the-art learning tasks. Unfortunately, while very powerful, Deep Learning is not well understood theoretically and in particular only recent results for the complexity of training deep neural networks have been obtained. In this work we show that large classes of deep neural networks with various architectures (e.g., DNNs, CNNs, Binary Neural Networks, and ResNets), activation functions (e.g., ReLUs and leaky ReLUs), and loss functions (e.g., Hinge loss, Euclidean loss, etc) can be trained to near optimality with desired target accuracy using linear programming in time that is exponential in the input data and parameter space dimension and polynomial in the size of the data set; improvements of the dependence in the input dimension are known to be unlikely assuming  $P \neq NP$ , and improving the dependence on the parameter space dimension remains open. In particular, we obtain polynomial time algorithms for training for a given fixed network architecture. Our work applies more broadly to empirical risk minimization.

minimization problems which allows us to generalize various previous results and obtain new complexity results for previously unstudied architectures in the proper learning setting.

\*\*\*\*\*

#### Cutting Down Training Memory by Re-forwarding

Jianwei Feng, Dong Huang

Deep Neural Networks (DNNs) require huge GPU memory when training on modern image/video databases. Unfortunately, the GPU memory as a hardware resource is always finite, which limits the image resolution, batch size, and learning rate that could be used for better DNN performance. In this paper, we propose a novel training approach, called Re-forwarding, that substantially reduces memory usage in training. Our approach automatically finds a subset of vertices in a DNN computation graph, and stores tensors only at these vertices during the first forward. During backward, extra local forwards (called the Re-forwarding process) are conducted to compute the missing tensors between the subset of vertices. The total memory cost becomes the sum of (1) the memory cost at the subset of vertices and (2) the maximum memory cost among local re-forwards. Re-forwarding trades training time overheads for memory and does not compromise any performance in testing. We propose theories and algorithms that achieve the optimal memory solutions for DNNs with either linear or arbitrary computation graphs. Experiments show that Re-forwarding cuts down up-to 80% of training memory on popular DNNs such as Alexnet, VGG, ResNet, Densenet and Inception net.

\*\*\*\*\*

#### Neural Network Cost Landscapes as Quantum States

Abdulah Fawaz, Sebastien Piat, Paul Klein, Peter Mountney, Simone Severini

Quantum computers promise significant advantages over classical computers for a number of different applications. We show that the complete loss function landscape of a neural network can be represented as the quantum state output by a quantum computer. We demonstrate this explicitly for a binary neural network and, further, show how a quantum computer can train the network by manipulating this state using a well-known algorithm known as quantum amplitude amplification. We further show that with minor adaptation, this method can also represent the meta-loss landscape of a number of neural network architectures simultaneously. We search this meta-loss landscape with the same method to simultaneously train and design a binary neural network.

\*\*\*\*\*

#### Déjà Vu: An Empirical Evaluation of the Memorization Properties of Convnets

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Hervé Jégou

Convolutional neural networks memorize part of their training data, which is why strategies such as data augmentation and drop-out are employed to mitigate overfitting. This paper considers the related question of "membership inference", where the goal is to determine if an image was used during training. We consider membership tests over either ensembles of samples or individual samples. First, we show how to detect if a dataset was used to train a model, and in particular whether some validation images were used at train time. Then, we introduce a new approach to infer membership when a few of the top layers are not available or have been fine-tuned, and show that lower layers still carry information about the training samples. To support our findings, we conduct large-scale experiments on Imagenet and subsets of YFCC-100M with modern architectures such as VGG and Resnet.

\*\*\*\*\*

#### On the Ineffectiveness of Variance Reduced Optimization for Deep Learning

Aaron Defazio

The application of stochastic variance reduction to optimization has shown remarkable recent theoretical and practical success. The applicability of these techniques to the hard non-convex optimization problems encountered during training of modern deep neural networks is an open problem. We show that naive application of the SVRG technique and related approaches fail, and explore why.

\*\*\*\*\*



## Online Learning for Supervised Dimension Reduction

Ning Zhang, Qiang Wu

Online learning has attracted great attention due to the increasing demand for systems that have the ability of learning and evolving. When the data to be processed is also high dimensional and dimension reduction is necessary for visualization or prediction enhancement, online dimension reduction will play an essential role. The purpose of this paper is to propose new online learning approaches for supervised dimension reduction. Our first algorithm is motivated by adapting the sliced inverse regression (SIR), a pioneer and effective algorithm for supervised dimension reduction, and making it implementable in an incremental manner. The new algorithm, called incremental sliced inverse regression (ISIR), is able to update the subspace of significant factors with intrinsic lower dimensionality fast and efficiently when new observations come in. We also refine the algorithm by using an overlapping technique and develop an incremental overlapping sliced inverse regression (IOSIR) algorithm. We verify the effectiveness and efficiency of both algorithms by simulations and real data applications.

\*\*\*\*\*

## Monge-Ampère Flow for Generative Modeling

Linfeng Zhang, Weinan E, Lei Wang

We present a deep generative model, named Monge-Ampère flow, which builds on continuous-time gradient flow arising from the Monge-Ampère equation in optimal transport theory. The generative map from the latent space to the data space follows a dynamical system, where a learnable potential function guides a compressible fluid to flow towards the target density distribution. Training of the model amounts to solving an optimal control problem. The Monge-Ampère flow has tractable likelihoods and supports efficient sampling and inference. One can easily impose symmetry constraints in the generative model by designing suitable scalar potential functions. We apply the approach to unsupervised density estimation of the MNIST dataset and variational calculation of the two-dimensional Ising model at the critical point. This approach brings insights and techniques from Monge-Ampère equation, optimal transport, and fluid dynamics into reversible flow-based generative models.

\*\*\*\*\*

## Unsupervised Image to Sequence Translation with Canvas-Drawer Networks

Kevin Frans, Chin-Yi Cheng

Encoding images as a series of high-level constructs, such as brush strokes or discrete shapes, can often be key to both human and machine understanding. In many cases, however, data is only available in pixel form. We present a method for generating images directly in a high-level domain (e.g. brush strokes), without the need for real pairwise data. Specifically, we train a "canvas" network to imitate the mapping of high-level constructs to pixels, followed by a high-level "drawing" network which is optimized through this mapping towards solving a desired image recreation or translation task. We successfully discover sequential vector representations of symbols, large sketches, and 3D objects, utilizing only pixel data. We display applications of our method in image segmentation, and present several ablation studies comparing various configurations.

\*\*\*\*\*

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

Alexia Jolicoeur-Martineau

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

r 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>.

\*\*\*\*\*

Overcoming catastrophic forgetting through weight consolidation and long-term memory

Shixian Wen, Laurent Itti

Sequential learning of multiple tasks in artificial neural networks using gradient descent leads to catastrophic forgetting, whereby previously learned knowledge is erased during learning of new, disjoint knowledge. Here, we propose a new approach to sequential learning which leverages the recent discovery of adversarial examples. We use adversarial subspaces from previous tasks to enable learning of new tasks with less interference. We apply our method to sequentially learning to classify digits 0, 1, 2 (task 1), 4, 5, 6, (task 2), and 7, 8, 9 (task 3) in MNIST (disjoint MNIST task). We compare and combine our Adversarial Direction (AD) method with the recently proposed Elastic Weight Consolidation (EWC) method for sequential learning. We train each task for 20 epochs, which yields good initial performance (99.24% correct task 1 performance). After training task 2, and then task 3, both plain gradient descent (PGD) and EWC largely forget task 1 (task 1 accuracy 32.95% for PGD and 41.02% for EWC), while our combined approach (AD+EWC) still achieves 94.53% correct on task 1. We obtain similar results with a much more difficult disjoint CIFAR10 task (70.10% initial task 1 performance, 67.73% after learning tasks 2 and 3 for AD+EWC, while PGD and EWC both fall to chance level). We confirm qualitatively similar results for EMNIST with 5 tasks and under 3 variants of our approach. Our results suggest that AD+EWC can provide better sequential learning performance than either PGD or EWC.

\*\*\*\*\*

Collapse of deep and narrow neural nets

Lu Lu, Yanhui Su, George Em Karniadakis

Recent theoretical work has demonstrated that deep neural networks have superior performance over shallow networks, but their training is more difficult, e.g., they suffer from the vanishing gradient problem. This problem can be typically resolved by the rectified linear unit (ReLU) activation. However, here we show that even for such activation, deep and narrow neural networks (NNs) will converge to erroneous mean or median states of the target function depending on the loss with high probability. Deep and narrow NNs are encountered in solving partial differential equations with high-order derivatives. We demonstrate this collapse of such NNs both numerically and theoretically, and provide estimates of the probability of collapse. We also construct a diagram of a safe region for designing NNs that avoid the collapse to erroneous states. Finally, we examine different ways of initialization and normalization that may avoid the collapse problem. Asymmetric initializations may reduce the probability of collapse but do not totally eliminate it.

\*\*\*\*\*

Stop memorizing: A data-dependent regularization framework for intrinsic pattern learning

Wei Zhu, Qiang Qiu, Bao Wang, Jianfeng Lu, Guillermo Sapiro, Ingrid Daubechies

Deep neural networks (DNNs) typically have enough capacity to fit random data by brute force even when conventional data-dependent regularizations focusing on the geometry of the features are imposed. We find out that the reason for this is the inconsistency between the enforced geometry and the standard softmax cross entropy loss. To resolve this, we propose a new framework for data-dependent DNN regularization, the Geometrically-Regularized-Self-Validating neural Networks (GRSVNet). During training, the geometry enforced on one batch of features is simultaneously validated on a separate batch using a validation loss consistent with the geometry. We study a particular case of GRSVNet, the Orthogonal-Low-rank Embedding (OLE)-GRSVNet, which is capable of producing highly discriminative features residing in orthogonal low-rank subspaces. Numerical experiments show that OLE-GRSVNet outperforms DNNs with conventional regularization when trained on real data. More importantly, unlike conventional DNNs, OLE-GRSVNet refuses to memorize random data or random labels, suggesting it only learns intrinsic patterns by reducing the memorizing capacity of the baseline DNN.

\*\*\*\*\*

SupportNet: solving catastrophic forgetting in class incremental learning with support data

Yu Li, Zhongxiao Li, Lizhong Ding, Yijie Pan, Chao Huang, Yuhui Hu, Wei Chen, Xin Gao

A plain well-trained deep learning model often does not have the ability to learn new knowledge without forgetting the previously learned knowledge, which is known as catastrophic forgetting. Here we propose a novel method, SupportNet, to efficiently and effectively solve the catastrophic forgetting problem in the class incremental learning scenario. SupportNet combines the strength of deep learning and support vector machine (SVM), where SVM is used to identify the support data from the old data, which are fed to the deep learning model together with the new data for further training so that the model can review the essential information of the old data when learning the new information. Two powerful consolidation regularizers are applied to stabilize the learned representation and ensure the robustness of the learned model. We validate our method with comprehensive experiments on various tasks, which show that SupportNet drastically outperforms the state-of-the-art incremental learning methods and even reaches similar performance as the deep learning model trained from scratch on both old and new data.

.

\*\*\*\*\*

A Kernel Random Matrix-Based Approach for Sparse PCA

Mohamed El Amine Seddik, Mohamed Tamaazousti, Romain Couillet

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  $n, p \rightarrow \infty$  with  $p/n \rightarrow c \in (0, \infty)$  and under Gaussian vector observations, we study kernel random matrices of the type  $f(\mathbf{M})$ , where  $f$  is a three-times continuously differentiable function applied entry-wise to the sample covariance matrix  $\mathbf{M}$  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.

\*\*\*\*\*

Neural Variational Inference For Embedding Knowledge Graphs

Alexander I. Cowen-Rivers, Pasquale Minervini

Recent advances in Neural Variational Inference allowed for a renaissance in latent variable models in a variety of domains involving high-dimensional data. In this paper, we introduce two generic Variational Inference frameworks for generative models of Knowledge Graphs; Latent Fact Model and Latent Information Model.

While traditional variational methods derive an analytical approximation for the intractable distribution over the latent variables, here we construct an inference network conditioned on the symbolic representation of entities and relation types in the Knowledge Graph, to provide the variational distributions. The new framework can create models able to discover underlying probabilistic semantics for the symbolic representation by utilising parameterisable distributions which

ch permit training by back-propagation in the context of neural variational inference, resulting in a highly-scalable method. Under a Bernoulli sampling framework, we provide an alternative justification for commonly used techniques in large-scale stochastic variational inference, which drastically reduces training time at a cost of an additional approximation to the variational lower bound. The generative frameworks are flexible enough to allow training under any prior distribution that permits a re-parametrisation trick, as well as under any scoring function that permits maximum likelihood estimation of the parameters. Experiment results display the potential and efficiency of this framework by improving upon multiple benchmarks with Gaussian prior representations. Code publicly available on Github.

\*\*\*\*\*

k-Nearest Neighbors by Means of Sequence to Sequence Deep Neural Networks and Memory Networks

Yiming Xu, Diego Klabjan

k-Nearest Neighbors is one of the most fundamental but effective classification models. In this paper, we propose two families of models built on a sequence to sequence model and a memory network model to mimic the k-Nearest Neighbors model, which generate a sequence of labels, a sequence of out-of-sample feature vectors and a final label for classification, and thus they could also function as oversamplers. We also propose 'out-of-core' versions of our models which assume that only a small portion of data can be loaded into memory. Computational experiments show that our models outperform k-Nearest Neighbors, a feed-forward neural network and a memory network, due to the fact that our models must produce additional output and not just the label. As an oversampler on imbalanced datasets, the sequence to sequence kNN model often outperforms Synthetic Minority Over-sampling Technique and Adaptive Synthetic Sampling.

\*\*\*\*\*

Clinical Risk: wavelet reconstruction networks for marked point processes

Jeremy C. Weiss

Timestamped sequences of events, pervasive in domains with data logs, e.g., health records, are often modeled as point processes with rate functions over time. Leading classical methods for risk scores such as Cox and Hawkes processes use such data but make strong assumptions about the shape and form of multivariate influences, resulting in time-to-event distributions irreflexive of many real world processes. Recent methods in point processes and recurrent neural networks capably model rate functions but may be complex and difficult to interrogate. Our work develops a high-performing, interrogable model. We introduce wavelet reconstruction networks, a multivariate point process with a sparse wavelet reconstruction kernel to model rate functions from marked, timestamped data. We show they achieve improved performance and interrogability over baselines in forecasting complications and scheduled care visits in patients with diabetes.

\*\*\*\*\*

Caveats for information bottleneck in deterministic scenarios

Artemy Kolchinsky, Brendan D. Tracey, Steven Van Kuyk

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",  $I(Y;T) - \beta I(X;T)$ . 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  $\beta$ ; (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 pr

oposal. 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

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.

\*\*\*\*\*

$\mathcal{A}^*$  sampling with probability matching

Yichi Zhou, Jun Zhu

Probabilistic methods often need to draw samples from a nontrivial distribution.

$\mathcal{A}^*$  sampling is a nice algorithm by building upon a top-down construction of a Gumbel process, where a large state space is divided into subsets and at each round  $\mathcal{A}^*$  sampling selects a subset to process. However, the selection rule depends on a bound function, which can be intractable. Moreover, we show that such a selection criterion can be inefficient. This paper aims to improve  $\mathcal{A}^*$  sampling by addressing these issues. To design a suitable selection rule, we apply `\emph{Probability Matching}`, a widely used method for decision making, to  $\mathcal{A}^*$  sampling. We provide insights into the relationship between  $\mathcal{A}^*$  sampling and probability matching by analyzing a nontrivial special case in which the state space is partitioned into two subsets. We show that in this case probability matching is optimal within a constant gap. Furthermore, as directly applying probability matching to  $\mathcal{A}^*$  sampling is time consuming, we design an approximate version based on Monte-Carlo estimators. We also present an efficient implementation by leveraging special properties of Gumbel distributions and well-designed balanced trees. Empirical results show that our method saves a significant amount of computational resources on suboptimal regions compared with  $\mathcal{A}^*$  sampling.

\*\*\*\*\*

Convergence Properties of Deep Neural Networks on Separable Data

Remi Tachet des Combes, Mohammad Pezeshki, Samira Shabanian, Aaron Courville, Yoshua Bengio

While a lot of progress has been made in recent years, the dynamics of learning in deep nonlinear neural networks remain to this day largely misunderstood. In this work, we study the case of binary classification and prove various properties of learning in such networks under strong assumptions such as linear separability of the data. Extending existing results from the linear case, we confirm empirical observations by proving that the classification error also follows a sigmoidal shape in nonlinear architectures. We show that given proper initialization, learning expounds parallel independent modes and that certain regions of parameter space might lead to failed training. We also demonstrate that input norm and features' frequency in the dataset lead to distinct convergence speeds which might shed some light on the generalization capabilities of deep neural networks. We provide a comparison between the dynamics of learning with cross-entropy and hinge losses, which could prove useful to understand recent progress in the tra

ining of generative adversarial networks. Finally, we identify a phenomenon that we baptize gradient starvation where the most frequent features in a dataset prevent the learning of other less frequent but equally informative features.

\*\*\*\*\*

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

David Berthelot\*, Colin Raffel\*, Aurko Roy, Ian Goodfellow

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

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.

\*\*\*\*\*

#### Excitation Dropout: Encouraging Plasticity in Deep Neural Networks

Andrea Zunino, Sarah Adel Bargal, Pietro Morerio, Jianming Zhang, Stan Sclaroff, Vittorio Murino

We propose a guided dropout regularizer for deep networks based on the evidence of a network prediction: the firing of neurons in specific paths. In this work, we utilize the evidence at each neuron to determine the probability of dropout, rather than dropping out neurons uniformly at random as in standard dropout. In essence, we dropout with higher probability those neurons which contribute more to decision making at training time. This approach penalizes high saliency neurons that are most relevant for model prediction, i.e. those having stronger evidence. By dropping such high-saliency neurons, the network is forced to learn alternative paths in order to maintain loss minimization, resulting in a plasticity-like behavior, a characteristic of human brains too. We demonstrate better generalization ability, an increased utilization of network neurons, and a higher resilience to network compression using several metrics over four image/video recognition benchmarks.

\*\*\*\*\*

#### Generating Liquid Simulations with Deformation-aware Neural Networks

Lukas Prantl, Boris Bonev, Nils Thuerey

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.

\*\*\*\*\*

## Stochastic Gradient Descent Learns State Equations with Nonlinear Activations

Samet Oymak

We study discrete time dynamical systems governed by the state equation  $h_{t+1} = \phi(Ah_t + Bu_t)$ . Here  $A, B$  are weight matrices,  $\phi$  is an activation function, and  $u_t$  is the input data. This relation is the backbone of recurrent neural networks (e.g. LSTMs) which have broad applications in sequential learning tasks. We utilize stochastic gradient descent to learn the weight matrices from a finite input/state trajectory  $(u_t, h_t)_{t=0}^{N-1}$ . We prove that SGD estimate linearly converges to the ground truth weights while using near-optimal sample size. Our results apply to increasing activations whose derivatives are bounded away from zero. The analysis is based on i) an SGD convergence result with nonlinear activations and ii) careful statistical characterization of the state vector. Numerical experiments verify the fast convergence of SGD on ReLU and leaky ReLU in consistency with our theory.

\*\*\*\*\*

## A Resizable Mini-batch Gradient Descent based on a Multi-Armed Bandit

Seong Jin Cho, Sunghun Kang, Chang D. Yoo

Determining the appropriate batch size for mini-batch gradient descent is always time consuming as it often relies on grid search. This paper considers a resizable mini-batch gradient descent (RMGD) algorithm based on a multi-armed bandit that achieves performance equivalent to that of best fixed batch-size. At each epoch, the RMGD samples a batch size according to a certain probability distribution proportional to a batch being successful in reducing the loss function. Sampling from this probability provides a mechanism for exploring different batch size and exploiting batch sizes with history of success. After obtaining the validation loss at each epoch with the sampled batch size, the probability distribution is updated to incorporate the effectiveness of the sampled batch size. Experimental results show that the RMGD achieves performance better than the best performing single batch size. It is surprising that the RMGD achieves better performance than grid search. Furthermore, it attains this performance in a shorter amount of time than grid search.

\*\*\*\*\*

## L2-Nonexpansive Neural Networks

Haifeng Qian, Mark N. Wegman

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.

perties.

\*\*\*\*\*

## Optimizing for Generalization in Machine Learning with Cross-Validation Gradients

Barratt, Shane, Sharma, Rishi

Cross-validation is the workhorse of modern applied statistics and machine learning, as it provides a principled framework for selecting the model that maximizes generalization performance. In this paper, we show that the cross-validation risk is differentiable with respect to the hyperparameters and training data for many common machine learning algorithms, including logistic regression, elastic-net regression, and support vector machines. Leveraging this property of differentiability, we propose a cross-validation gradient method (CVGM) for hyperparameter optimization. Our method enables efficient optimization in high-dimensional hyperparameter spaces of the cross-validation risk, the best surrogate of the true generalization ability of our learning algorithm.

\*\*\*\*\*

## Discriminative Active Learning

Daniel Gissin, Shai Shalev-Shwartz

We propose a new batch mode active learning algorithm designed for neural networks and large query batch sizes. The method, Discriminative Active Learning (DAL), poses active learning as a binary classification task, attempting to choose examples to label in such a way as to make the labeled set and the unlabeled pool indistinguishable. Experimenting on image classification tasks, we empirically show our method to be on par with state of the art methods in medium and large query batch sizes, while being simple to implement and also extend to other domains besides classification tasks. Our experiments also show that none of the state of the art methods of today are clearly better than uncertainty sampling, negating some of the reported results in the recent literature.

\*\*\*\*\*

## Improving On-policy Learning with Statistical Reward Accumulation

Yubin Deng, Ke Yu, Dahua Lin, Xiaoou Tang, Chen Change Loy

Deep reinforcement learning has obtained significant breakthroughs in recent years. Most methods in deep-RL achieve good results via the maximization of the reward signal provided by the environment, typically in the form of discounted cumulative returns. Such reward signals represent the immediate feedback of a particular action performed by an agent. However, tasks with sparse reward signals are still challenging to on-policy methods. In this paper, we introduce an effective characterization of past reward statistics (which can be seen as long-term feedback signals) to supplement this immediate reward feedback. In particular, value functions are learned with multi-critics supervision, enabling complex value functions to be more easily approximated in on-policy learning, even when the reward signals are sparse. We also introduce a novel exploration mechanism called 'hot-wiring' that can give a boost to seemingly trapped agents. We demonstrate the effectiveness of our advantage actor multi-critic (A2MC) method across the discrete domains in Atari games as well as continuous domains in the MuJoCo environments. A video demo is provided at <https://youtu.be/zBmpf3Yz8tc> and source codes will be made available upon paper acceptance.

\*\*\*\*\*

## Quantization for Rapid Deployment of Deep Neural Networks

Jun Haeng Lee, Sangwon Ha, Saerom Choi, Won-Jo Lee, Seungwon Lee

This paper aims at rapid deployment of the state-of-the-art deep neural networks (DNNs) to energy efficient accelerators without time-consuming fine tuning or the availability of the full datasets. Converting DNNs in full precision to limited precision is essential in taking advantage of the accelerators with reduced memory footprint and computation power. However, such a task is not trivial since it often requires the full training and validation datasets for profiling the network statistics and fine tuning the networks to recover the accuracy lost after quantization. To address these issues, we propose a simple method recognizing channel-level distribution to reduce the quantization-induced accuracy loss and minimize the required image samples for profiling. We evaluated our method on e



leven networks trained on the ImageNet classification benchmark and a network trained on the Pascal VOC object detection benchmark. The results prove that the networks can be quantized into 8-bit integer precision without fine tuning.

\*\*\*\*\*

Explicit Information Placement on Latent Variables using Auxiliary Generative Modelling Task

Nat Diloekthanakul, Nick Pawlowski, Murray Shanahan

Deep latent variable models, such as variational autoencoders, have been successfully used to disentangle factors of variation in image datasets. The structure of the representations learned by such models is usually observed after training and iteratively refined by tuning the network architecture and loss function. Here we propose a method that can explicitly place information into a specific subset of the latent variables. We demonstrate the use of the method in a task of disentangling global structure from local features in images. One subset of the latent variables is encouraged to represent local features through an auxiliary modelling task. In this auxiliary task, the global structure of an image is destroyed by dividing it into pixel patches which are then randomly shuffled. The full set of latent variables is trained to model the original data, obliging the remainder of the latent representation to model the global structure. We demonstrate that this approach successfully disentangles the latent variables for global structure from local structure by observing the generative samples of SVHN and CIFAR10. We also cluster the disentangled global structure of SVHN and found that the emerging clusters represent meaningful groups of global structures - including digit identities and the number of digits presence. Finally, we discuss the problem of evaluating the clustering accuracy when ground truth categories are not expressive enough.

\*\*\*\*\*

Provable Guarantees on Learning Hierarchical Generative Models with Deep CNNs

Eran Malach, Shai Shalev-Shwartz

Learning deep networks is computationally hard in the general case. To show any positive theoretical results, one must make assumptions on the data distribution. Current theoretical works often make assumptions that are very far from describing real data, like sampling from Gaussian distribution or linear separability of the data. We describe an algorithm that learns convolutional neural network, assuming the data is sampled from a deep generative model that generates images level by level,

where lower resolution images correspond to latent semantic classes. We analyze the convergence rate of our algorithm assuming the data is indeed generated according to this model (as well as

additional assumptions). While we do not pretend to claim that the assumptions are realistic for natural images, we do believe that they capture some true properties of real data. Furthermore, we show that on CIFAR-10, the algorithm we analyze achieves results in the same ballpark with vanilla convolutional neural networks that are trained with SGD.

\*\*\*\*\*

Efficient Dictionary Learning with Gradient Descent

Dar Gilboa, Sam Buchanan, John Wright

Randomly initialized first-order optimization algorithms are the method of choice for solving many high-dimensional nonconvex problems in machine learning, yet general theoretical guarantees cannot rule out convergence to critical points of poor objective value. For some highly structured nonconvex problems however, the success of gradient descent can be understood by studying the geometry of the objective. We study one such problem -- complete orthogonal dictionary learning, and provide convergence guarantees for randomly initialized gradient descent to the neighborhood of a global optimum. The resulting rates scale as low order polynomials in the dimension even though the objective possesses an exponential number of saddle points. This efficient convergence can be viewed as a consequence of negative curvature normal to the stable manifolds associated with saddle points, and we provide evidence that this feature is shared by other nonconvex problems of importance as well.

\*\*\*\*\*

Accidental exploration through value predictors

Tomasz Kisielewski, Damian Leśniak, Maia Pasek

Infinite length of trajectories is an almost universal assumption in the theoretical foundations of reinforcement learning. In practice learning occurs on finite trajectories. In this paper we examine a specific result of this disparity, namely a strong bias of the time-bounded Every-visit Monte Carlo value estimator. This manifests as a vastly different learning dynamic for algorithms that use value predictors, including encouraging or discouraging exploration.

We investigate these claims theoretically for a one dimensional random walk, and empirically on a number of simple environments. We use GAE as an algorithm involving a value predictor and evolution strategies as a reference point.

\*\*\*\*\*

Deep, Skinny Neural Networks are not Universal Approximators

Jesse Johnson

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.

\*\*\*\*\*

Metric-Optimized Example Weights

Sen Zhao, Mahdi Milani Fard, Maya Gupta

Real-world machine learning applications often have complex test metrics, and may have training and test data that follow different distributions. We propose addressing these issues by using a weighted loss function with a standard convex loss, but with weights on the training examples that are learned to optimize the test metric of interest on the validation set. These metric-optimized example weights can be learned for any test metric, including black box losses and customized metrics for specific applications. We illustrate the performance of our proposal with public benchmark datasets and real-world applications with domain shift and custom loss functions that balance multiple objectives, impose fairness policies, and are non-convex and non-decomposable.

\*\*\*\*\*

Laplacian Smoothing Gradient Descent

Stanley J. Osher, Bao Wang, Penghang Yin, Xiyang Luo, Minh Pham, Alex T. Lin

We propose a class of very simple modifications of gradient descent and stochastic gradient descent. We show that when applied to a large variety of machine learning problems, ranging from softmax regression to deep neural nets, the proposed surrogates can dramatically reduce the variance and improve the generalization accuracy. The methods only involve multiplying the usual (stochastic) gradient by the inverse of a positive definite matrix coming from the discrete Laplacian or its high order generalizations. The theory of Hamilton-Jacobi partial differential equations demonstrates that the implicit version of new algorithm is almost the same as doing gradient descent on a new function which (i) has the same global minima as the original function and (ii) is "more convex". We show that optimization algorithms with these surrogates converge uniformly in the discrete Sobolev  $H_{\sigma^p}$  sense and reduce the optimality gap for convex optimization problems. We implement our algorithm into both PyTorch and Tensorflow platforms which only involves changing of a few lines of code. The code will be available on Github.

\*\*\*\*\*

Reduced-Gate Convolutional LSTM Design Using Predictive Coding for Next-Frame Video Prediction

Nelly Elsayed, Anthony S. Maida, Magdy Bayoumi

Spatiotemporal sequence prediction is an important problem in deep learning. We study next-frame video prediction using a deep-learning-based predictive coding framework that uses convolutional, long short-term memory (convLSTM) modules. We introduce a novel reduced-gate convolutional LSTM architecture. Our reduced-gate model achieves better next-frame prediction accuracy than the original

convolutional LSTM while using a smaller parameter budget, thereby reducing training time. We tested our reduced gate modules within a predictive coding architecture

on the moving MNIST and KITTI datasets. We found that our reduced-gate model has a significant reduction of approximately 40 percent of the total number of training parameters and training time in comparison with the standard LSTM model which makes it attractive for hardware implementation especially on small devices.

\*\*\*\*\*

#### Targeted Adversarial Examples for Black Box Audio Systems

Rohan Taori, Amog Kamsetty, Brenton Chu, Nikita Vemuri

The application of deep recurrent networks to audio transcription has led to impressive gains in automatic speech recognition (ASR) systems. Many have demonstrated that small adversarial perturbations can fool deep neural networks into incorrectly predicting a specified target with high confidence. Current work on fooling ASR systems have focused on white-box attacks, in which the model architecture and parameters are known. In this paper, we adopt a black-box approach to adversarial generation, combining the approaches of both genetic algorithms and gradient estimation to solve the task. We achieve a 89.25% targeted attack similarity after 3000 generations while maintaining 94.6% audio file similarity.

\*\*\*\*\*

#### Large Scale Graph Learning From Smooth Signals

Vassilis Kalofolias, Nathanaël Perraudin

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.

\*\*\*\*\*

#### Discovering Low-Precision Networks Close to Full-Precision Networks for Efficient Embedded Inference

Jeffrey L. McKinstry, Steven K. Esser, Rathinakumar Appuswamy, Deepika Bablani, John V. Arthur, Izzet B. Yildiz, Dharmendra S. Modha

To realize the promise of ubiquitous embedded deep network inference, it is essential to seek limits of energy and area efficiency. To this end, low-precision networks offer tremendous promise because both energy and area scale down quadratically with the reduction in precision. Here, for the first time, we demonstrate ResNet-18, ResNet-34, ResNet-50, ResNet-152, Inception-v3, densenet-161, and VGG-16bn networks on the ImageNet classification benchmark that, at 8-bit precision exceed the accuracy of the full-precision baseline networks after one epoch of finetuning, thereby leveraging the availability of pretrained models.

We also demonstrate ResNet-18, ResNet-34, and ResNet-50 4-bit models that match the accuracy of the full-precision baseline networks -- the highest scores to date. Surprisingly, the weights of the low-precision networks are very close (in cosine similarity) to the weights of the corresponding baseline networks, making training from scratch unnecessary.

We find that gradient noise due to quantization during training increases with  $r$

duced precision, and seek ways to overcome this noise. The number of iterations required by stochastic gradient descent to achieve a given training error is related to the square of (a) the distance of the initial solution from the final plus (b) the maximum variance of the gradient estimates. By drawing inspiration from this observation, we (a) reduce solution distance by starting with pretrained fp32 precision baseline networks and fine-tuning, and (b) combat noise introduced by quantizing weights and activations during training, by using larger batches along with matched learning rate annealing. Sensitivity analysis indicates that these techniques, coupled with proper activation function range calibration, offer a promising heuristic to discover low-precision networks, if they exist, close to fp32 precision baseline networks.

\*\*\*\*\*

Weak contraction mapping and optimization

Siwei Luo

The weak contraction mapping is a self mapping that the range is always a subset of the domain, which admits a unique fixed-point. The iteration of weak contraction mapping is a Cauchy sequence that yields the unique fixed-point. A gradient-free optimization method as an application of weak contraction mapping is proposed to achieve global minimum convergence. The optimization method is robust to local minima and initial point position.

\*\*\*\*\*

IEA: Inner Ensemble Average within a convolutional neural network

Abduallah Mohamed,Xinrui Hua,Xianda Zhou,Christian Claudel

Ensemble learning is a method of combining multiple trained models to improve model accuracy. We propose the usage of such methods, specifically ensemble average, inside Convolutional Neural Network (CNN) architectures by replacing the single convolutional layers with Inner Average Ensembles (IEA) of multiple convolutional layers. Empirical results on different benchmarking datasets show that CNN models using IEA outperform those with regular convolutional layers and advances the state of art. A visual and a similarity score analysis of the features generated from IEA explains why it boosts the model performance.

\*\*\*\*\*

Explaining Adversarial Examples with Knowledge Representation

Xingyu Zhou,Tengyu Ma,Huahong Zhang

Adversarial examples are modified samples that preserve original image structures but deviate classifiers. Researchers have put efforts into developing methods for generating adversarial examples and finding out origins. Past research put much attention on decision boundary changes caused by these methods. This paper, in contrast, discusses the origin of adversarial examples from a more underlying knowledge representation point of view. Human beings can learn and classify prototypes as well as transformations of objects. While neural networks store learned knowledge in a more hybrid way of combining all prototypes and transformations as a whole distribution. Hybrid storage may lead to lower distances between different classes so that small modifications can mislead the classifier. A one-step distribution imitation method is designed to imitate distribution of the nearest different class neighbor. Experiments show that simply by imitating distributions from a training set without any knowledge of the classifier can still lead to obvious impacts on classification results from deep networks. It also implies that adversarial examples can be in more forms than small perturbations. Potential ways of alleviating adversarial examples are discussed from the representation point of view. The first path is to change the encoding of data sent to the training step. Training data that are more prototypical can help seize more robust and accurate structural knowledge. The second path requires constructing learning frameworks with improved representations.

\*\*\*\*\*

Learning Graph Representations by Dendrograms

Thomas Bonald,Bertrand Charpentier

Hierarchical clustering is a common approach to analysing the multi-scale structure of graphs observed in practice.

We propose a novel metric for assessing the quality of a hierarchical clustering. This metric reflects the ability to reconstruct the graph from the dendrogram encoding the hierarchy. The best representation of the graph for this metric in turn yields a novel hierarchical clustering algorithm. Experiments on both real and synthetic data illustrate the efficiency of the approach.

\*\*\*\*\*

Large-scale classification of structured objects using a CRF with deep class embedding

Eran Goldman, Jacob Goldberger

This paper presents a novel deep learning architecture for classifying structured objects in ultrafine-grained datasets, where classes may not be clearly distinguishable by their appearance but rather by their context. We model sequences of images as linear-chain CRFs, and jointly learn the parameters from both local visual features and neighboring class information. The visual features are learned by convolutional layers, whereas class-structure information is reparametrized by factorizing the CRF pairwise potential matrix. This forms a context-based semantic similarity space, learned alongside the visual similarities, and dramatically increases the learning capacity of contextual information. This new parametrization, however, forms a highly nonlinear objective function which is challenging to optimize. To overcome this, we develop a novel surrogate likelihood which allows for a local likelihood approximation of the original CRF with integrated batch-normalization. This model overcomes the difficulties of existing CRF methods to learn the contextual relationships thoroughly when there is a large number of classes and the data is sparse. The performance of the proposed method is illustrated on a huge dataset that contains images of retail-store product displays, and shows significantly improved results compared to linear CRF parametrization, unnormalized likelihood optimization, and RNN modeling.

\*\*\*\*\*

A unified theory of adaptive stochastic gradient descent as Bayesian filtering  
Laurence Aitchison

We formulate stochastic gradient descent (SGD) as a novel factorised Bayesian filtering problem, in which each parameter is inferred separately, conditioned on the corresponding backpropagated gradient. Inference in this setting naturally gives rise to BRMSprop and BAdam: Bayesian variants of RMSprop and Adam. Remarkably, the Bayesian approach recovers many features of state-of-the-art adaptive SGD methods, including amongst others root-mean-square normalization, Nesterov acceleration and AdamW. As such, the Bayesian approach provides one explanation for the empirical effectiveness of state-of-the-art adaptive SGD algorithms. Empirically comparing BRMSprop and BAdam with naive RMSprop and Adam on MNIST, we find that Bayesian methods have the potential to considerably reduce test loss and classification error.

\*\*\*\*\*

The Forward-Backward Embedding of Directed Graphs

Thomas Bonald, Nathan De Lara

We introduce a novel embedding of directed graphs derived from the singular value decomposition (SVD) of the normalized adjacency matrix. Specifically, we show that, after proper normalization of the singular vectors, the distances between vectors in the embedding space are proportional to the mean commute times between the corresponding nodes by a forward-backward random walk in the graph, which follows the edges alternately in forward and backward directions. In particular, two nodes having many common successors in the graph tend to be represented by close vectors in the embedding space. More formally, we prove that our representation of the graph is equivalent to the spectral embedding of some co-citation graph, where nodes are linked with respect to their common set of successors in the original graph. The interest of our approach is that it does not require to build this co-citation graph, which is typically much denser than the original graph. Experiments on real datasets show the efficiency of the approach.

\*\*\*\*\*

## Variation Network: Learning High-level Attributes for Controlled Input Manipulation

Gaëtan Hadjeres

This paper presents the Variation Network (VarNet), a generative model providing means to manipulate the high-level attributes of a given input. The originality of our approach is that VarNet is not only capable of handling pre-defined attributes but can also learn the relevant attributes of the dataset by itself. These two settings can be easily combined which makes VarNet applicable for a wide variety of tasks. Further, VarNet has a sound probabilistic interpretation which grants us with a novel way to navigate in the latent spaces as well as means to control how the attributes are learned. We demonstrate experimentally that this model is capable of performing interesting input manipulation and that the learned attributes are relevant and interpretable.

\*\*\*\*\*

## A Synaptic Neural Network and Synapse Learning

Chang Li

A Synaptic Neural Network (SynaNN) consists of synapses and neurons. Inspired by the synapse research of neuroscience, we built a synapse model with a nonlinear synapse function of excitatory and inhibitory channel probabilities. Introduced the concept of surprisal space and constructed a commutative diagram, we proved that the inhibitory probability function  $-\log(1-\exp(-x))$  in surprisal space is the topologically conjugate function of the inhibitory complementary probability  $1-x$  in probability space. Furthermore, we found that the derivative of the synapse over the parameter in the surprisal space is equal to the negative Bose-Einstein distribution. In addition, we constructed a fully connected synapse graph (tensor) as a synapse block of a synaptic neural network. Moreover, we proved the gradient formula of a cross-entropy loss function over parameters, so synapse learning can work with the gradient descent and backpropagation algorithms. In the proof-of-concept experiment, we performed an MNIST training and testing on the MLP model with synapse network as hidden layers.

\*\*\*\*\*

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

Xiuyuan Cheng, Qiang Qiu, Robert Calderbank, Guillermo Sapiro

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

The high computational and parameter complexity of neural networks makes their training very slow and difficult to deploy on energy and storage-constrained computing systems. Many network complexity reduction techniques have been proposed including fixed-point implementation. However, a systematic approach for designing 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 feedforw

ard 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 suitable 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.

\*\*\*\*\*

Deep Neuroevolution: Genetic Algorithms are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning

Felipe Petroski Such, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O. Stanley, Jeff Clune

Deep artificial neural networks (DNNs) are typically trained via gradient-based learning algorithms, namely backpropagation.

Evolution strategies (ES) can rival backprop-based algorithms such as Q-learning and policy gradients on challenging deep reinforcement learning (RL) problems. However, ES can be considered a gradient-based algorithm because it performs stochastic gradient descent via an operation similar to a finite-difference approximation of the gradient.

That raises the question of whether non-gradient-based evolutionary algorithms can work at DNN scales.

Here we demonstrate they can: we evolve the weights of a DNN with a simple, gradient-free, population-based genetic algorithm (GA) and it performs well on hard deep RL problems, including Atari and humanoid locomotion. The Deep GA successfully evolves networks with over four million free parameters, the largest neural networks ever evolved with a traditional evolutionary algorithm. These results (1) expand our sense of the scale at which GAs can operate, (2) suggest intriguingly that in some cases following the gradient is not the best choice for optimizing performance, and (3) make immediately available the multitude of neuroevolution techniques that improve performance. We demonstrate the latter by showing that combining DNNs with novelty search, which encourages exploration on tasks with deceptive or sparse reward functions, can solve a high-dimensional problem on which reward-maximizing algorithms (e.g., DQN, A3C, ES, and the GA) fail. Additionally, the Deep GA is faster than ES, A3C, and DQN (it can train Atari in  $\sim 4$  hours on one workstation or  $\sim 1$  hour distributed on 720 cores), and enables a state-of-the-art, up to 10,000-fold compact encoding technique.

\*\*\*\*\*

Playing the Game of Universal Adversarial Perturbations

Julien Perolet, Mateusz Malinowski, Bilal Piot, Olivier Pietquin

We study the problem of learning classifiers robust to universal adversarial perturbations. While prior work approaches this problem via robust optimization, adversarial training, or input transformation, we instead phrase it as a two-player zero-sum game. In this new formulation, both players simultaneously play the same game, where one player chooses a classifier that minimizes a classification loss whilst the other player creates an adversarial perturbation that increases the same loss when applied to every sample in the training set.

By observing that performing a classification (respectively creating adversarial samples) is the best response to the other player, we propose a novel extension of a game-theoretic algorithm, namely fictitious play, to the domain of training robust classifiers. Finally, we empirically show the robustness and versatility of our approach in two defence scenarios where universal attacks are performed on several image classification datasets -- CIFAR10, CIFAR100 and ImageNet.

\*\*\*\*\*

Lipschitz regularized Deep Neural Networks generalize

Adam M. Oberman, Jeff Calder

We show that if the usual training loss is augmented by a Lipschitz regularization term, then the networks generalize. We prove generalization by first establishing

shing a stronger convergence result, along with a rate of convergence. A second result resolves a question posed in Zhang et al. (2016): how can a model distinguish between the case of clean labels, and randomized labels? Our answer is that Lipschitz regularization using the Lipschitz constant of the clean data makes this distinction. In this case, the model learns a different function which we hypothesize correctly fails to learn the dirty labels.

\*\*\*\*\*

Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourages convex latent distributions

Tim Sainburg, Marvin Thielk, Brad Thielman, Benjamin Migliori, Timothy Gentner

We present a neural network architecture based upon the Autoencoder (AE) and Generative Adversarial Network (GAN) that promotes a convex latent distribution by training adversarially on latent space interpolations. By using an AE as both the generator and discriminator of a GAN, we pass a pixel-wise error function across the discriminator, yielding an AE which produces sharp samples that match both high- and low-level features of the original images. Samples generated from interpolations between data in latent space remain within the distribution of real data as trained by the discriminator, and therefore preserve realistic resemblances to the network inputs.

\*\*\*\*\*

Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, Jack Xin

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-modified 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.

\*\*\*\*\*

Deep Perm-Set Net: Learn to predict sets with unknown permutation and cardinality using deep neural networks

S. Hamid Reza Tofighi, Roman Kaskman, Farbod T. Motlagh, Qinfeng Shi, Daniel Cremers, Laura Leal-Taixé, Ian Reid

Many real-world problems, e.g. object detection, have outputs that are naturally expressed as sets of entities. This creates a challenge for traditional deep neural networks which naturally deal with structured outputs such as vectors, matrices or tensors. We present a novel approach for learning to predict sets with unknown permutation and cardinality using deep neural networks. Specifically, in our formulation we incorporate the permutation as unobservable variable and estimate its distribution during the learning process using alternating optimization. We demonstrate the validity of this new formulation on two relevant vision problems: object detection, for which our formulation outperforms state-of-the-art detectors such as Faster R-CNN and YOLO, and a complex CAPTCHA test, where we observe that, surprisingly, our set based network acquired the ability of mimicking arithmetics without any rules being coded.

\*\*\*\*\*



## Count-Based Exploration with the Successor Representation

Marlos C. Machado, Marc G. Bellemare, Michael Bowling

The problem of exploration in reinforcement learning is well-understood in the tabular case and many sample-efficient algorithms are known. Nevertheless, it is often unclear how the algorithms in the tabular setting can be extended to tasks with large state-spaces where generalization is required. Recent promising developments generally depend on problem-specific density models or handcrafted features. In this paper we introduce a simple approach for exploration that allows us to develop theoretically justified algorithms in the tabular case but that also give us intuitions for new algorithms applicable to settings where function approximation is required. Our approach and its underlying theory is based on the substochastic successor representation, a concept we develop here. While the traditional successor representation is a representation that defines state generalization by the similarity of successor states, the substochastic successor representation is also able to implicitly count the number of times each state (or feature) has been observed. This extension connects two until now disjoint areas of research. We show in traditional tabular domains (RiverSwim and SixArms) that our algorithm empirically performs as well as other sample-efficient algorithms. We then describe a deep reinforcement learning algorithm inspired by these ideas and show that it matches the performance of recent pseudo-count-based methods in hard exploration Atari 2600 games.

\*\*\*\*\*

## Multiple Encoder-Decoders Net for Lane Detection

Yuetong Du, Xiaodong Gu, Junqin Liu, Liwen He

For semantic image segmentation and lane detection, nets with a single spatial pyramid structure or encoder-decoder structure are usually exploited. Convolutional neural networks (CNNs) show great results on both high-level and low-level features representations, however, the capability has not been fully embodied for lane detection task. In especial, it's still a challenge for model-based lane detection to combine the multi-scale context with a pixel-level accuracy because of the weak visual appearance and strong prior information. In this paper, we propose a novel network for lane detection, the three main contributions are as follows. First, we employ multiple encoder-decoders module in end-to-end ways and show the promising results for lane detection. Second, we analysis different configurations of multiple encoder-decoders nets. Third, we make our attempts to rethink the evaluation methods of lane detection for the limitation of the popular methods based on IoU.

\*\*\*\*\*

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

Chiyu Max Jiang, Dequan Wang, Jingwei Huang, Philip Marcus, Matthias Niessner

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

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.

\*\*\*\*\*

Towards Language Agnostic Universal Representations

Armen Aghajanyan,Xia Song,Saurabh Tiwary

When a bilingual student learns to solve word problems in math, we expect the student to be able to solve these problem in both languages the student is fluent in, even if the math lessons were only taught in one language. However, current representations in machine learning are language dependent. In this work, we present a method to decouple the language from the problem by learning language agnostic representations and therefore allowing training a model in one language and applying to a different one in a zero shot fashion. We learn these representations by taking inspiration from linguistics, specifically the Universal Grammar hypothesis and learn universal latent representations that are language agnostic (Chomsky, 2014; Montague, 1970). We demonstrate the capabilities of these representations by showing that the models trained on a single language using language agnostic representations achieve very similar accuracies in other languages.

\*\*\*\*\*

Precision Highway for Ultra Low-precision Quantization

Eunhyeok Park,Dongyoung Kim,Sungjoo Yoo,Peter Vajda

Quantization of a neural network has an inherent problem called accumulated quantization error, which is the key obstacle towards ultra-low precision, e.g., 2- or 3-bit precision. To resolve this problem, we propose precision highway, which forms an end-to-end high-precision information flow while performing the ultra-low-precision computation. First, we describe how the precision highway reduce the accumulated quantization error in both convolutional and recurrent neural networks. We also provide the quantitative analysis of the benefit of precision highway and evaluate the overhead on the state-of-the-art hardware accelerator. In the experiments, our proposed method outperforms the best existing quantization methods while offering 3-bit weight/activation quantization with no accuracy loss and 2-bit quantization with a 2.45 % top-1 accuracy loss in ResNet-50. We also report that the proposed method significantly outperforms the existing method in the 2-bit quantization of an LSTM for language modeling.

\*\*\*\*\*

Traditional and Heavy Tailed Self Regularization in Neural Network Models

Charles H. Martin,Michael W. Mahoney

Random Matrix Theory (RMT) is applied to analyze the weight matrices of Deep Neural Networks (DNNs), including both production quality, pre-trained models such as AlexNet and Inception, and smaller models trained from scratch, such as LeNet

5 and a miniature-AlexNet. Empirical and theoretical results clearly indicate that the empirical spectral density (ESD) of DNN layer matrices displays signatures of traditionally-regularized statistical models, even in the absence of exogenously specifying traditional forms of regularization, such as Dropout or Weight Norm constraints. Building on recent results in RMT, most notably its extension to Universality classes of Heavy-Tailed matrices, we develop a theory to identify 5+1 Phases of Training, corresponding to increasing amounts of Implicit Self-Regularization. For smaller and/or older DNNs, this Implicit Self-Regularization is like traditional Tikhonov regularization, in that there is a "size scale" separating signal from noise. For state-of-the-art DNNs, however, we identify a novel form of Heavy-Tailed Self-Regularization, similar to the self-organization seen in the statistical physics of disordered systems. This implicit Self-Regularization can depend strongly on the many knobs of the training process. By exploiting the generalization gap phenomena, we demonstrate that we can cause a small model to exhibit all 5+1 phases of training simply by changing the batch size.

\*\*\*\*\*

Filter Training and Maximum Response: Classification via Discerning

Lei Gu

This report introduces a training and recognition scheme, in which classification is realized via class-wise discerning. Trained with datasets whose labels are randomly shuffled except for one class of interest, a neural network learns class-wise parameter values, and remolds itself from a feature sorter into feature filters, each of which discerns objects belonging to one of the classes only. Classification of an input can be inferred from the maximum response of the filters. A multiple check with multiple versions of filters can diminish fluctuation and yields better performance. This scheme of discerning, maximum response and multiple check is a method of general viability to improve performance of feedforward networks, and the filter training itself is a promising feature abstraction procedure. In contrast to the direct sorting, the scheme mimics the classification process mediated by a series of one component picking.

\*\*\*\*\*

The meaning of "most" for visual question answering models

Alexander Kuhnle, Ann Copestake

The correct interpretation of quantifier statements in the context of a visual scene requires non-trivial inference mechanisms. For the example of "most", we discuss two strategies which rely on fundamentally different cognitive concepts. Our aim is to identify what strategy deep learning models for visual question answering learn when trained on such questions. To this end, we carefully design data to replicate experiments from psycholinguistics where the same question was investigated for humans. Focusing on the FiLM visual question answering model, our experiments indicate that a form of approximate number system emerges whose performance declines with more difficult scenes as predicted by Weber's law. Moreover, we identify confounding factors, like spatial arrangement of the scene, which impede the effectiveness of this system.

\*\*\*\*\*

Diminishing Batch Normalization

Yintai Ma, Diego Klabjan

In this paper, we propose a generalization of the BN algorithm, diminishing batch normalization (DBN), where we update the BN parameters in a diminishing moving average way. Batch normalization (BN) is very effective in accelerating the convergence of a neural network training phase that it has become a common practice.

Our proposed DBN algorithm remains the overall structure of the original BN algorithm while introduces a weighted averaging update to some trainable parameters.

We provide an analysis of the convergence of the DBN algorithm that converges to a stationary point with respect to trainable parameters. Our analysis can be easily generalized for original BN algorithm by setting some parameters to constant. To the best knowledge of authors, this analysis is the first of its kind for

convergence with Batch Normalization introduced. We analyze a two-layer model with arbitrary activation function.

The primary challenge of the analysis is the fact that some parameters are updated by gradient while others are not.

The convergence analysis applies to any activation function that satisfies our common assumptions.

For the analysis, we also show the sufficient and necessary conditions for the stepsizes and diminishing weights to ensure the convergence.

In the numerical experiments, we use more complex models with more layers and ReLU activation. We observe that DBN outperforms the original BN algorithm on ImageNet, MNIST, NI and CIFAR-10 datasets with reasonable complex FNN and CNN models.

\*\*\*\*\*

FEED: Feature-level Ensemble Effect for knowledge Distillation

SeongUk Park, Nojun Kwak

This paper proposes a versatile and powerful training algorithm named Feature-level Ensemble Effect for knowledge Distillation (FEED), which is inspired by the work of factor transfer. The factor transfer is one of the knowledge transfer methods that improves the performance of a student network with a strong teacher network. It transfers the knowledge of a teacher in the feature map level using high-capacity teacher network, and our training algorithm FEED is an extension of it. FEED aims to transfer ensemble knowledge, using either multiple teachers in parallel or multiple training sequences. Adapting the peer-teaching framework, we introduce a couple of training algorithms that transfer ensemble knowledge to the student at the feature map level, both of which help the student network find more generalized solutions in the parameter space. Experimental results on CIFAR-100 and ImageNet show that our method, FEED, has clear performance enhancements, without introducing any additional parameters or computations at test time.

\*\*\*\*\*

Invariance and Inverse Stability under ReLU

Jens Behrmann, Sören Dittmer, Pascal Fernsel, Peter Maass

We flip the usual approach to study invariance and robustness of neural networks by considering the non-uniqueness and instability of the inverse mapping. We provide theoretical and numerical results on the inverse of ReLU-layers. First, we derive a necessary and sufficient condition on the existence of invariance that provides a geometric interpretation. Next, we move to robustness via analyzing local effects on the inverse. To conclude, we show how this reverse point of view not only provides insights into key effects, but also enables to view adversarial examples from different perspectives.

\*\*\*\*\*

Amortized Context Vector Inference for Sequence-to-Sequence Networks

Sotirios Chatzis, Kyriacos Toliass, Aristotelis Charalampous

Neural attention (NA) has become a key component of sequence-to-sequence models that yield state-of-the-art performance in as hard tasks as abstractive document summarization (ADS), machine translation (MT), and video captioning (VC). NA mechanisms perform inference of context vectors; these constitute weighted sums of deterministic input sequence encodings, adaptively sourced over long temporal horizons. Inspired from recent work in the field of amortized variational inference (AVI), in this work we consider treating the context vectors generated by soft-attention (SA) models as latent variables, with approximate finite mixture model posteriors inferred via AVI. We posit that this formulation may yield stronger generalization capacity, in line with the outcomes of existing applications of AVI to deep networks. To illustrate our method, we implement it and experimentally evaluate it considering challenging ADS, VC, and MT benchmarks. This way, we exhibit its improved effectiveness over state-of-the-art alternatives.

\*\*\*\*\*

CHEMICAL NAMES STANDARDIZATION USING NEURAL SEQUENCE TO SEQUENCE MODEL

Junlang Zhan, Hai Zhao

Chemical information extraction is to convert chemical knowledge in text into true chemical database, which is a text processing task heavily relying on chemical

1 compound name identification and standardization. Once a systematic name for a chemical compound is given, it will naturally and much simply convert the name into the eventually required molecular formula. However, for many chemical substances, they have been shown in many other names besides their systematic names which poses a great challenge for this task. In this paper, we propose a framework to do the auto standardization from the non-systematic names to the corresponding systematic names by using the spelling error correction, byte pair encoding tokenization and neural sequence to sequence model. Our framework is trained end to end and is fully data-driven. Our standardization accuracy on the test dataset achieves 54.04% which has a great improvement compared to previous state-of-the-art result.

\*\*\*\*\*

#### Classification from Positive, Unlabeled and Biased Negative Data

Yu-Guan Hsieh, Gang Niu, Masashi Sugiyama

Positive-unlabeled (PU) learning addresses the problem of learning a binary classifier from positive (P) and unlabeled (U) data. It is often applied to situations where negative (N) data are difficult to be fully labeled. However, collecting a non-representative N set that contains only a small portion of all possible N data can be much easier in many practical situations. This paper studies a novel classification framework which incorporates such biased N (bN) data in PU learning. The fact that the training N data are biased also makes our work very different from those of standard semi-supervised learning. We provide an empirical risk minimization-based method to address this PUBN classification problem. Our approach can be regarded as a variant of traditional example-reweighting algorithms, with the weight of each example computed through a preliminary step that draws inspiration from PU learning. We also derive an estimation error bound for the proposed method. Experimental results demonstrate the effectiveness of our algorithm in not only PUBN learning scenarios but also ordinary PU learning scenarios on several benchmark datasets.

\*\*\*\*\*

#### Interpreting Layered Neural Networks via Hierarchical Modular Representation

Chihiro Watanabe

Interpreting the prediction mechanism of complex models is currently one of the most important tasks in the machine learning field, especially with layered neural networks, which have achieved high predictive performance with various practical data sets. To reveal the global structure of a trained neural network in an interpretable way, a series of clustering methods have been proposed, which decompose the units into clusters according to the similarity of their inference roles. The main problems in these studies were that (1) we have no prior knowledge about the optimal resolution for the decomposition, or the appropriate number of clusters, and (2) there was no method with which to acquire knowledge about whether the outputs of each cluster have a positive or negative correlation with the input and output dimension values.

In this paper, to solve these problems, we propose a method for obtaining a hierarchical modular representation of a layered neural network. The application of a hierarchical clustering method to a trained network reveals a tree-structured relationship among hidden layer units, based on their feature vectors defined by their correlation with the input and output dimension values.

\*\*\*\*\*

#### ATTENTION INCORPORATE NETWORK: A NETWORK CAN ADAPT VARIOUS DATA SIZE

Liangbo He, Hao Sun

In traditional neural networks for image processing, the inputs of the neural networks should be the same size such as  $224 \times 224 \times 3$ . But how can we train the neural net model with different input size? A common way to do is image deformation which accompany a problem of information loss (e.g. image crop or wrap). In this paper we propose a new network structure called Attention Incorporate Network (AIN). It solve the problem of different size of input images and extract the key features of the inputs by attention mechanism, pay different attention depends on the importance of the features not rely on the data size. Experimentally, AIN achieve a higher accuracy, better convergence comparing to the same size of other

network structure.

\*\*\*\*\*

The Universal Approximation Power of Finite-Width Deep ReLU Networks

Dmytro Perekhrestenko, Philipp Grohs, Dennis Elbrächter, Helmut Bölcskei

We show that finite-width deep ReLU neural networks yield rate-distortion optimal approximation (Bölcskei et al., 2018) of a wide class of functions, including polynomials, windowed sinusoidal functions, one-dimensional oscillatory textures, and the Weierstrass function, a fractal function which is continuous but nowhere differentiable. Together with the recently established universal approximation result for affine function systems (Bölcskei et al., 2018), this demonstrates that deep neural networks approximate vastly different signal structures generated by the affine group, the Weyl-Heisenberg group, or through warping, and even certain fractals, all with approximation error decaying exponentially in the number of neurons. We also prove that in the approximation of sufficiently smooth functions finite-width deep networks require strictly fewer neurons than finite-depth wide networks.

\*\*\*\*\*

CONTROLLING COVARIATE SHIFT USING EQUILIBRIUM NORMALIZATION OF WEIGHTS

Aaron Defazio

We introduce a new normalization technique that exhibits the fast convergence properties of batch normalization using a transformation of layer weights instead of layer outputs. The proposed technique keeps the contribution of positive and negative weights to the layer output in equilibrium. We validate our method on a set of standard benchmarks including CIFAR-10/100, SVHN and ILSVRC 2012 ImageNet.

\*\*\*\*\*

Deep Curiosity Search: Intra-Life Exploration Can Improve Performance on Challenging Deep Reinforcement Learning Problems

Christopher Stanton, Jeff Clune

Traditional exploration methods in reinforcement learning (RL) require agents to perform random actions to find rewards. But these approaches struggle on sparse-reward domains like Montezuma's Revenge where the probability that any random action sequence leads to reward is extremely low. Recent algorithms have performed well on such tasks by encouraging agents to visit new states or perform new actions in relation to all prior training episodes (which we call across-training novelty). But such algorithms do not consider whether an agent exhibits intra-life novelty: doing something new within the current episode, regardless of whether those behaviors have been performed in previous episodes. We hypothesize that across-training novelty might discourage agents from revisiting initially non-rewarding states that could become important stepping stones later in training—a problem remedied by encouraging intra-life novelty. We introduce Curiosity Search for deep reinforcement learning, or Deep Curiosity Search (DeepCS), which encourages intra-life exploration by rewarding agents for visiting as many different states as possible within each episode, and show that DeepCS matches the performance of current state-of-the-art methods on Montezuma's Revenge. We further show that DeepCS improves exploration on Amidar, Freeway, Gravitar, and Tutankham (many of which are hard exploration games). Surprisingly, DeepCS also doubles A2C performance on Seaquest, a game we would not have expected to benefit from intra-life exploration because the arena is small and already easily navigated by naive exploration techniques. In one run, DeepCS achieves a maximum training score of 80,000 points on Seaquest—higher than any methods other than Ape-X. The strong performance of DeepCS on these sparse- and dense-reward tasks suggests that encouraging intra-life novelty is an interesting, new approach for improving performance in Deep RL and motivates further research into hybridizing across-training and intra-life exploration methods.

\*\*\*\*\*

(Unconstrained) Beam Search is Sensitive to Large Search Discrepancies

Eldan Cohen, J. Christopher Beck

Beam search is the most popular inference algorithm for decoding neural sequence models. Unlike greedy search, beam search allows for a non-greedy local decision

ns that can potentially lead to a sequence with a higher overall probability. However, previous work found that the performance of beam search tends to degrade with large beam widths. In this work, we perform an empirical study of the behavior of the beam search algorithm across three sequence synthesis tasks. We find that increasing the beam width leads to sequences that are disproportionately based on early and highly non-greedy decisions. These sequences typically include a very low probability token that is followed by a sequence of tokens with higher (conditional) probability leading to an overall higher probability sequence. However, as beam width increases, such sequences are more likely to have a lower evaluation score. Based on our empirical analysis we propose to constrain the beam search from taking highly non-greedy decisions early in the search. We evaluate two methods to constrain the search and show that constrained beam search effectively eliminates the problem of beam search degradation and in some cases even leads to higher evaluation scores. Our results generalize and improve upon previous observations on copies and training set predictions.

\*\*\*\*\*

Training for Faster Adversarial Robustness Verification via Inducing ReLU Stability

Kai Y. Xiao, Vincent Tjeng, Nur Muhammad (Mahi) Shafiullah, Aleksander Madry

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.

\*\*\*\*\*

RESIDUAL NETWORKS CLASSIFY INPUTS BASED ON THEIR NEURAL TRANSIENT DYNAMICS

Fereshteh Lagzi

In this study, we analyze the input-output behavior of residual networks from a dynamical system point of view by disentangling the residual dynamics from the output activities before the classification stage. For a network with simple skip connections between every successive layer, and for logistic activation function, and shared weights between layers, we show analytically that there is a cooperation and competition dynamics between residuals corresponding to each input dimension.

Interpreting these kind of networks as nonlinear filters, the steady state value of the residuals in the case of attractor networks are indicative of the common features between different input dimensions that the network has observed during training, and has encoded in those components. In cases where residuals do not converge to an attractor state, their internal dynamics are separable for each input class, and the network can reliably approximate the output. We bring analytical and

empirical evidence that residual networks classify inputs based on the integration of the transient dynamics of the residuals, and will show how the network responds to input perturbations. We compare the network dynamics for a ResNet and a Multi-Layer Perceptron and show that the internal dynamics, and the noise evolution are fundamentally different in these networks, and ResNets are more robust to noisy inputs. Based on these findings, we also develop a new method to adjust the depth for residual networks during training. As it turns out, after pruning the depth of a ResNet using this algorithm, the network is still capable of classifying inputs with a high accuracy.

\*\*\*\*\*

Unsupervised classification into unknown number of classes

Sungyeob Han, Daeyoung Kim, Jungwoo Lee

We propose a novel unsupervised classification method based on graph Laplacian. Unlike the widely used classification method, this architecture does not require the labels of data and the number of classes. Our key idea is to introduce an approximate linear map and a spectral clustering theory on the dimension reduced spaces into generative adversarial networks. Inspired by the human visual recognition system, the proposed framework can classify and also generate images as the human brains do. We build an approximate linear connector network  $\mathcal{C}$  analogous to the cerebral cortex, between the discriminator  $\mathcal{D}$  and the generator  $\mathcal{G}$ . The connector network allows us to estimate the unknown number of classes. Estimating the number of classes is one of the challenging researches in the unsupervised learning, especially in spectral clustering. The proposed method can also classify the images by using the estimated number of classes. Therefore, we define our method as an unsupervised classification method.

\*\*\*\*\*

#### Smoothing the Geometry of Probabilistic Box Embeddings

Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, Andrew McCallum

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.

\*\*\*\*\*