Machine Learning Meets Public Policy: What to Expect and How to Cope

Invited Talk | Tue Dec 4th 08:30 -- 09:20 AM @ Rooms 220 CDE

Edward W Felten

AI and Machine Learning are already having a big impact on the world. Policymakers have noticed, and they are starting to formulate laws and regulations, and to convene conversations, about how society will govern the development of these technologies. This talk will give an overview of how policymakers deal with new technologies, how the process might develop in the case of AI/ML, and why constructive engagement with the policy process will lead to better outcomes for the field, for governments, and for society.

Coffee Break

Break | Tue Dec 4th 09:40 -- 10:05 AM @

—

On Neuronal Capacity

Oral | Tue Dec 4th 10:05 -- 10:20 AM @ Room 220 CD

Pierre Baldi · Roman Vershynin

We define the capacity of a learning machine to be the logarithm of the number (or volume) of the functions it can implement. We review known results, and derive new results, estimating the capacity of several neuronal models: linear and polynomial threshold gates, linear and polynomial threshold gates with constrained weights (binary weights, positive weights), and ReLU neurons. We also derive capacity estimates and bounds for fully recurrent networks and layered feedforward networks.

On the Dimensionality of Word Embedding

Oral | Tue Dec 4th 10:05 -- 10:20 AM @ Room 220 E

Zi Yin · Yuanyuan Shen

In this paper, we provide a theoretical understanding of word embedding and its dimensionality.

Motivated by the unitary-invariance of word embedding, we propose the Pairwise Inner Product (PIP) loss, a novel metric on the dissimilarity between word embeddings. Using techniques from matrix perturbation theory, we reveal a fundamental bias-variance trade-off in dimensionality selection for word embeddings. This bias-variance trade-off sheds light on many empirical observations which were previously unexplained, for example the existence of an optimal dimensionality. Moreover, new insights and discoveries, like when and how word embeddings are robust to over-fitting, are revealed. By optimizing over the bias-variance trade-off of the PIP loss, we can explicitly answer the open question of dimensionality selection for word embedding.

Phase Retrieval Under a Generative Prior

Oral | Tue Dec 4th 10:05 -- 10:20 AM @ Room 517 CD

Paul Hand · Oscar Leong · Vlad Voroninski

We introduce a novel deep-learning inspired formulation of the \textit{phase retrieval problem}, which asks to recover a signal $y_0 \in \R^n$ from m quadratic observations, under structural assumptions on the underlying signal. As is common in many imaging problems, previous methodologies have considered natural signals as being sparse with respect to a known basis, resulting in the decision to enforce a generic sparsity prior. However, these methods for phase retrieval have encountered possibly fundamental limitations, as no computationally efficient algorithm for sparse phase retrieval has been proven to succeed with fewer than $O(k^2\log n)$ generic measurements, which is larger than the theoretical optimum of $O(k \log n)$. In this paper, we sidestep this issue by considering a prior that a natural signal is in the range of a generative neural network $G : \R^k \rightarrow \R^n$. We introduce an empirical risk formulation that has favorable global geometry for gradient methods, as soon as $m = O(k)$, under the model of a multilayer fully-connected neural network with random weights. Specifically, we show that there exists a descent direction outside of a small neighborhood around the true k-dimensional latent code and a negative multiple thereof. This formulation for structured phase retrieval thus benefits from two effects: generative priors can more tightly represent natural signals than sparsity priors, and this empirical risk formulation can exploit those generative priors at an information theoretically optimal sample complexity, unlike for a sparsity prior. We corroborate these results with experiments showing that exploiting generative models in phase retrieval tasks outperforms both sparse and general phase retrieval methods.

Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data

Spotlight | Tue Dec 4th 10:20 -- 10:25 AM @ Room 220 CD

Yuanzhi Li · Yingyu Liang

Neural networks have many successful applications, while much less theoretical understanding has been gained. Towards bridging this gap, we study the problem of learning a two-layer overparameterized ReLU neural network for multi-class classification via stochastic gradient descent (SGD) from random initialization. In the overparameterized setting, when the data comes from mixtures of well-separated distributions, we prove that SGD learns a network with a small generalization error, albeit the network has enough capacity to fit arbitrary labels. Furthermore, the analysis provides interesting insights into several aspects of learning neural networks and can be verified based on empirical studies on synthetic data and on the MNIST dataset.

Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces

Spotlight | Tue Dec 4th 10:20 -- 10:25 AM @ Room 220 E

Yu-An Chung · Wei-Hung Weng · Schrasing Tong · James Glass

Recent research has shown that word embedding spaces learned from text corpora of different languages can be aligned without any parallel data supervision. Inspired by the success in unsupervised cross-lingual word embeddings, in this paper we target learning a cross-modal alignment between the embedding spaces of speech and text learned from corpora of their respective modalities in an unsupervised fashion. The proposed framework learns the individual speech and text embedding spaces, and attempts to align the two spaces via adversarial training, followed by a refinement procedure. We show how our framework could be used to perform the tasks of spoken word classification and translation, and the experimental results on these two tasks demonstrate that the performance of our unsupervised alignment approach is comparable to its supervised counterpart. Our framework is especially useful for developing automatic speech recognition (ASR) and speech-to-text translation systems for low- or zero-resource languages, which have little parallel audio-text data for training modern supervised ASR and speech-to-text translation models, but account for the majority of the languages spoken across the world.

Global Geometry of Multichannel Sparse Blind Deconvolution on the Sphere

Spotlight | Tue Dec 4th 10:20 -- 10:25 AM @ Room 517 CD

Yanjun Li · Yoram Bresler

Multichannel blind deconvolution is the problem of recovering an unknown signal f and multiple unknown channels x_i from convolutional measurements $y_i=x_i \circledast f$ ($i=1,2,\dots,N$). We consider the case where the x_i's are sparse, and convolution with f is invertible. Our nonconvex optimization formulation solves for a filter h on the unit sphere that produces sparse output $y_i\circledast h$. Under some technical assumptions, we show that all local minima of the objective function correspond to the inverse filter of f up to an inherent sign and shift ambiguity, and all saddle points have strictly negative curvatures. This geometric structure allows successful recovery of f and x_i using a simple manifold gradient descent algorithm with random initialization. Our theoretical findings are complemented by numerical experiments, which demonstrate superior performance of the proposed approach over the previous methods.

Size-Noise Tradeoffs in Generative Networks

Spotlight | Tue Dec 4th 10:25 -- 10:30 AM @ Room 220 CD

Bolton Bailey · Matus Telgarsky

This paper investigates the ability of generative networks to convert their input noise distributions into other distributions. Firstly, we demonstrate a construction that allows ReLU networks to increase the dimensionality of their noise distribution by implementing a ``space-filling'' function based on iterated tent maps. We show this construction is optimal by analyzing the number of affine pieces in functions computed by multivariate ReLU networks. Secondly, we provide efficient ways (using polylog$(1/\epsilon)$ nodes) for networks to pass between univariate uniform and normal distributions, using a Taylor series approximation and a binary search gadget for computing function inverses. Lastly, we indicate how high dimensional distributions can be efficiently transformed into low dimensional distributions.

Diffusion Maps for Textual Network Embedding

Spotlight | Tue Dec 4th 10:25 -- 10:30 AM @ Room 220 E

Xinyuan Zhang · Yitong Li · Dinghan Shen · Lawrence Carin

Textual network embedding leverages rich text information associated with the network to learn low-dimensional vectorial representations of vertices. Rather than using typical natural language processing (NLP) approaches, recent research exploits the relationship of texts on the same edge to graphically embed text. However, these models neglect to measure the complete level of connectivity between any two texts in the graph. We present diffusion maps for textual network embedding (DMTE), integrating global structural information of the graph to capture the semantic relatedness between texts, with a diffusion-convolution operation applied on the text inputs. In addition, a new objective function is designed to efficiently preserve the high-order proximity using the graph diffusion. Experimental results show that the proposed approach outperforms state-of-the-art methods on the vertex-classification and link-prediction tasks.

Theoretical Linear Convergence of Unfolded ISTA and Its Practical Weights and Thresholds

Spotlight | Tue Dec 4th 10:25 -- 10:30 AM @ Room 517 CD

Xiaohan Chen · Jialin Liu · Zhangyang Wang · Wotao Yin

In recent years, unfolding iterative algorithms as neural networks has become an empirical success in solving sparse recovery problems. However, its theoretical understanding is still immature, which prevents us from fully utilizing the power of neural networks. In this work, we study unfolded ISTA (Iterative Shrinkage Thresholding Algorithm) for sparse signal recovery. We introduce a weight structure that is necessary for asymptotic convergence to the true sparse signal. With this structure, unfolded ISTA can attain a linear convergence, which is better than the sublinear convergence of ISTA/FISTA in general cases. Furthermore, we propose to incorporate thresholding in the network to perform support selection, which is easy to implement and able to boost the convergence rate both theoretically and empirically. Extensive simulations, including sparse vector recovery and a compressive sensing experiment on real image data, corroborate our theoretical results and demonstrate their practical usefulness. We have made our codes publicly available: https://github.com/xchen-tamu/linear-lista-cpss.

Dendritic cortical microcircuits approximate the backpropagation algorithm

Oral | Tue Dec 4th 10:30 -- 10:45 AM @ Room 220 CD

João Sacramento · Rui Ponte Costa · Yoshua Bengio · Walter Senn

Deep learning has seen remarkable developments over the last years, many of them inspired by neuroscience. However, the main learning mechanism behind these advances – error backpropagation – appears to be at odds with neurobiology. Here, we introduce a multilayer neuronal network model with simplified dendritic compartments in which error-driven synaptic plasticity adapts the network towards a global desired output. In contrast to previous work our model does not require separate phases and synaptic learning is driven by local dendritic prediction errors continuously in time. Such errors originate at apical dendrites and occur due to a mismatch between predictive input from lateral interneurons and activity from actual top-down feedback. Through the use of simple dendritic compartments and different cell-types our model can represent both error and normal activity within a pyramidal neuron. We demonstrate the learning capabilities of the model in regression and classification tasks, and show analytically that it approximates the error backpropagation algorithm. Moreover, our framework is consistent with recent observations of learning between brain areas and the architecture of cortical microcircuits. Overall, we introduce a novel view of learning on dendritic cortical circuits and on how the brain may solve the long-standing synaptic credit assignment problem.

A Retrieve-and-Edit Framework for Predicting Structured Outputs

Oral | Tue Dec 4th 10:30 -- 10:45 AM @ Room 220 E

Tatsunori Hashimoto · Kelvin Guu · Yonatan Oren · Percy Liang

For the task of generating complex outputs such as source code, editing existing outputs can be easier than generating complex outputs from scratch. With this motivation, we propose an approach that first retrieves a training example based on the input (e.g., natural language description) and then edits it to the desired output (e.g., code). Our contribution is a computationally efficient method for learning a retrieval model that embeds the input in a task-dependent way without relying on a hand-crafted metric or incurring the expense of jointly training the retriever with the editor. Our retrieve-and-edit framework can be applied on top of any base model. We show that on a new autocomplete task for GitHub Python code and the Hearthstone cards benchmark, retrieve-and-edit significantly boosts the performance of a vanilla sequence-to-sequence model on both tasks.

Spectral Filtering for General Linear Dynamical Systems

Oral | Tue Dec 4th 10:30 -- 10:45 AM @ Room 517 CD

Elad Hazan · HOLDEN LEE · Karan Singh · Cyril Zhang · Yi Zhang

We give a polynomial-time algorithm for learning latent-state linear dynamical systems without system identification, and without assumptions on the spectral radius of the system's transition matrix. The algorithm extends the recently introduced technique of spectral filtering, previously applied only to systems with a symmetric transition matrix, using a novel convex relaxation to allow for the efficient identification of phases.

TensorFlow Dance - Learning to Dance via Machine Learning

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510

Yaz Santissi · Jonathan DEKHTIAR

Deep learning to improve quality control in pharmaceutical manufacturing

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510 ABCD #D5

Michael Sass Hansen · Sebastian Brandes Kraaijenzank

This demo shows how deep learning can be applied in pharmaceutical industry, specifically for the reducing rejection rates of non-defect products in drug manufacturing. Advancements in convolutional neural networks for classification and variational autoencoders for anomaly detection have generated such impressive results over the past couple of years that the technology is now starting to become mature enough to be useful in the real world. Many drug manufacturers rely on highly manual, expensive processes for running their quality control operations and, until now, they haven't had a technological alternative advanced enough for being able to optimize this part of their manufacturing pipeline. Deep learning is a true game changer in this industry and being able to increase efficiency in the production of drugs leads potential huge price reductions, making modern medicine available to more people in need -- especially among low-income groups. This demo shows how these advantages can be obtained, as we will bring a professional CVT machine (capable of inspecting up to 600 cartridges or vials per minute) fitted with a chain of neural networks who run in real time to analyze the products and make decisions on whether to release or reject the products that pass by. Attendees will be able to interact with the underlying models through an easy-to-use interface that allows for retraining of models based on new datasets as well as deployment of the models. The goal of the demo is to leave attendees with the impression that neural nets are indeed ready to be deployed into highly regulated industries with the purpose of making a positive difference for all of us.

Ruuh: A Deep Learning Based Conversational Social Agent

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510 ABCD #D3

Puneet Agrawal · Manoj Kumar Chinnakotla · Sonam Damani · Meghana Joshi · Kedhar Nath Narahari · Khyatti Gupta · Nitya Raviprakash · Umang Gupta · Ankush Chatterjee · Abhishek Mathur · Sneha Magapu

Dialogue systems and conversational agents are becoming increasingly popular in the modern society but building an agent capable of holding intelligent conversation with its users is a challenging problem for artificial intelligence. In this demo, we demonstrate a deep learning based conversational social agent called “Ruuh” (facebook.com/Ruuh) designed by a team at Microsoft India to converse on a wide range of topics. Ruuh needs to think beyond the utilitarian notion of merely generating “relevant” responses and meet a wider range of user social needs, like expressing happiness when user's favorite team wins, sharing a cute comment on showing the pictures of the user's pet and so on. The agent also needs to detect and respond to abusive language, sensitive topics and trolling behavior of the users. Our agent has interacted with over 2 million real world users till date which has generated over 150 million user conversations to date.

A Hands-free Natural User Interface (NUI) for AR/VR Head-Mounted Displays Exploiting Wearer’s Facial Gestures

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510 ABCD #D7

Jaekwang Cha · Shiho Kim · Jinhyuk Kim

The demonstration presents interactions between head mounted display (HMD) worn user and augmented reality (AR) environment with our state-of-the-art hands-free user interface (UI) device which catches user’s facial gesture as an input signal of the UI. AR systems used in complex environment, such as surgery or works in dangerous environment, require a hands-free UI because they must continue to use their hands during operation. Moreover, hands-free UI helps improve the user experience (UX), not only in such a complex environment but also in common usage of AR and virtual reality (VR). Even though demands on interface device for HMD environment, there have not been such optimized interface yet like keyboard and mouse for PC or touch interface for smartphone. The objective of our demo is to present attendees a hands-free AR UI experience and to introduce attendees to benefits of using hands-free interface when using AR HMD environment. In the demo, attendee can deliver commands to the system through the wink gesture instead of using today’s common HMD input interface such as hand-held remote controller or HMD buttons which interferes user immerse on HMD environment. The wink acts like mouse click in our demonstration presented AR world. The facial gestures of user are automatically mapped to commands through deep neural networks. The proposed UI system is very unique and appropriate to develop various natural user interface (NUI) for AR/VR HMD environment because the sensing mechanism does not interfere user and allows user to hands-free.

Reproducing Machine Learning Research on Binder

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510

Jessica Forde · Tim Head · Chris Holdgraf · M Pacer · Félix-Antoine Fortin · Fernando Perez

Full author list is: Jessica Zosa Forde Matthias Bussonnier Félix-Antoine Fortin Brian Granger Tim Head Chris Holdgraf Paul Ivanov Kyle Kelley Fernando Perez M Pacer Yuvi Panda Gladys Nalvarte Min Ragan-Kelley Zach Sailer Steven Silvester Erik Sundell Carol Willing Researchers have encouraged the machine learning community to produce reproducible, complete pipelines for code. Binder is an open-source service that lets users share interactive, reproducible science. It uses standard configuration files in software engineering to create interactive versions of research that exist on sites like GitHub with minimal additional effort. By leveraging tools such as Kubernetes, it manages the technical complexity around creating containers to capture a repository and its dependencies, generating user sessions, and providing public URLs to share the built images with others. It combines two open-source projects within the Jupyter ecosystem: repo2docker and JupyterHub. repo2docker builds the Docker image of the git repository specified by the user, installs dependencies, and provides various front-ends to explore the image. JupyterHub then spawns and serves instances of these built images using Kubernetes to scale as needed. Our free public deployment, mybinder.org, features over 3,000 repos on topics such LIGO's gravational waves, textbooks on Kalman Filters, and open-source libraries such as PyMC3. As of September 2018, it serves an average of 8,000 users per day and has served as many as 22,000 a given day. Our demonstration shares a Binder deployment that features machine learning research papers from GitHub.

A model-agnostic web interface for interactive music composition by inpainting

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510 ABCD #D8

Gaëtan Hadjeres · Théis Bazin · Ashis Pati

We present a web-based interface that allows users to compose symbolic music in an interactive way using generative models for music. We strongly believe that such models only reveal their potential when used by artists and creators. While generative models for music have been around for a while, the conception of interactive interfaces designed for music creators is only burgeoning. We contribute to this emerging area by providing a general web interface for many music generation models so that researchers in the domain can easily test and promote their works. We hope that the present work will contribute in making A.I.-assisted composition accessible to a wider audience, from non musicians to professional musicians. This work is a concrete application of using music inpainting as a creative tool and could additionally be of interest to researchers in the domain for testing and evaluating their models. We show how we can use this system (generative model + interface) using different inpainting algorithms in an actual music production environment. The key elements of novelty are: (a) easy-to-use and intuitive interface for users, (b) easy-to-plug interface for researchers allowing them to explore the potential of their music generation algorithms, (c) web-based and model-agnostic framework, (d) integration of existing music inpainting algorithms, (e) novel inpainting algorithm for folk music, (f) novel paradigms for A.I.-assisted music composition and live performance, (g) integration in professional music production environments.

TextWorld: A Learning Environment for Text-based Games

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510 ABCD #D10

Marc-Alexandre Côté · Wendy Tay · Eric Yuan

Text-based games (e.g. Zork, Colossal Cave) are complex, interactive simulations in which text describes the game state and players make progress by entering text commands. They are fertile ground for language-focused machine learning research. In addition to language understanding, successful play requires skills like long-term memory and planning, exploration (trial and error), and common sense. This demonstration is about TextWorld, a Python-based learning environment for text-based games. TextWorld can be used to play existing games, as the ALE does for Atari games. However, the real novelty is that TextWorld can generate new text-based games with desired complexity. Its generative mechanisms give precise control over the difficulty, scope, and language of constructed games, and can therefore be used to study generalization and transfer learning.

Deep Neural Networks Running Onboard Anki’s Robot, Vector

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510 ABCD #D1

Lorenzo Riano · Andrew Stein · Mark Palatucci

In August Anki unveiled Vector, a home robot focused on personality and character. Vector is a palm-sized bot packed with unprecedented functionality in a very computationally constrained package. Among other capabilities, he uses deep neural networks to recognize elements of interest in the world, like people, hands, and other objects. At NIPS, we will discuss how we designed and tested the neural network architectures, the unique constraints that we had to face, and the solutions we developed. Since our network will have to run on hundred of thousands of robots worldwide, we had to develop unique metrics and testing methodologies to ensure that it provides the right data to various components that depend on it. We will describe how we limited the network footprint by employing quantization and pruning, and generally running neural networks on a constrained CPU. We will also show how perception is integrated into the bigger behavioral system to create a robot that is compelling and fun to interact with.

Game for Detecting Backdoor Attacks on Deep Neural Networks using Activation Clustering

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510 ABCD #D4

Casey Dugan · Werner Geyer · Aabhas Sharma · Ingrid Lange · Dustin Ramsey Torres · Bryant Chen · Nathalie Baracaldo Angel · Heiko Ludwig

NIPS Demo Submission_2.pdf

A machine learning environment to determine novel malaria policies

Demonstration | Tue Dec 4th 10:45 AM -- 07:30 PM @ Room 510 ABCD #D9

Oliver Bent · Sekou L Remy · Nelson Bore

The research and development of new tools and strategies in the fight against malaria, already uses resources, data and computation spread across innumerable institutions and individuals. Whether this is towards an objective such as drug discovery or informing intervention policy, they present common requirements. Such threads may be interwoven to achieve common goals towards malaria eradication. This unifying influence may be the technology of Artificial Intelligence (AI), helping to tie together different efforts, necessitating Novel Exploration Techniques for scientific discovery and an Infrastructure for Research at Scale.

Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #1

Xenia Miscouridou · Francois Caron · Yee Whye Teh

We propose a novel class of network models for temporal dyadic interaction data. Our objective is to capture important features often observed in social interactions: sparsity, degree heterogeneity, community structure and reciprocity. We use mutually-exciting Hawkes processes to model the interactions between each (directed) pair of individuals. The intensity of each process allows interactions to arise as responses to opposite interactions (reciprocity), or due to shared interests between individuals (community structure). For sparsity and degree heterogeneity, we build the non time dependent part of the intensity function on compound random measures following Todeschini et al., 2016. We conduct experiments on real-world temporal interaction data and show that the proposed model outperforms competing approaches for link prediction, and leads to interpretable parameters.

The Lingering of Gradients: How to Reuse Gradients Over Time

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #2

Zeyuan Allen-Zhu · David Simchi-Levi · Xinshang Wang

Classically, the time complexity of a first-order method is estimated by its number of gradient computations. In this paper, we study a more refined complexity by taking into account the ``lingering'' of gradients: once a gradient is computed at x_k, the additional time to compute gradients at $x_{k+1},x_{k+2},\dots$ may be reduced.

We show how this improves the running time of gradient descent and SVRG. For instance, if the "additional time'' scales linearly with respect to the traveled distance, then the "convergence rate'' of gradient descent can be improved from $1/T$ to $\exp(-T^{1/3})$. On the empirical side, we solve a hypothetical revenue management problem on the Yahoo! Front Page Today Module application with 4.6m users to 10^{-6} error (or 10^{-12} dual error) using 6 passes of the dataset.

Quadratic Decomposable Submodular Function Minimization

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #3

Pan Li · Niao He · Olgica Milenkovic

We introduce a new convex optimization problem, termed quadratic decomposable submodular function minimization. The problem is closely related to decomposable submodular function minimization and arises in many learning on graphs and hypergraphs settings, such as graph-based semi-supervised learning and PageRank. We approach the problem via a new dual strategy and describe an objective that may be optimized via random coordinate descent (RCD) methods and projections onto cones. We also establish the linear convergence rate of the RCD algorithm and develop efficient projection algorithms with provable performance guarantees. Numerical experiments in semi-supervised learning on hypergraphs confirm the efficiency of the proposed algorithm and demonstrate the significant improvements in prediction accuracy with respect to state-of-the-art methods.

On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #4

Lénaïc Chizat · Francis Bach

Many tasks in machine learning and signal processing can be solved by minimizing a convex function of a measure. This includes sparse spikes deconvolution or training a neural network with a single hidden layer. For these problems, we study a simple minimization method: the unknown measure is discretized into a mixture of particles and a continuous-time gradient descent is performed on their weights and positions. This is an idealization of the usual way to train neural networks with a large hidden layer. We show that, when initialized correctly and in the many-particle limit, this gradient flow, although non-convex, converges to global minimizers. The proof involves Wasserstein gradient flows, a by-product of optimal transport theory. Numerical experiments show that this asymptotic behavior is already at play for a reasonable number of particles, even in high dimension.

Leveraging the Exact Likelihood of Deep Latent Variable Models

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #5

Pierre-Alexandre Mattei · Jes Frellsen

Deep latent variable models (DLVMs) combine the approximation abilities of deep neural networks and the statistical foundations of generative models. Variational methods are commonly used for inference; however, the exact likelihood of these models has been largely overlooked. The purpose of this work is to study the general properties of this quantity and to show how they can be leveraged in practice. We focus on important inferential problems that rely on the likelihood: estimation and missing data imputation. First, we investigate maximum likelihood estimation for DLVMs: in particular, we show that most unconstrained models used for continuous data have an unbounded likelihood function. This problematic behaviour is demonstrated to be a source of mode collapse. We also show how to ensure the existence of maximum likelihood estimates, and draw useful connections with nonparametric mixture models. Finally, we describe an algorithm for missing data imputation using the exact conditional likelihood of a DLVM. On several data sets, our algorithm consistently and significantly outperforms the usual imputation scheme used for DLVMs.

DVAE#: Discrete Variational Autoencoders with Relaxed Boltzmann Priors

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #6

Arash Vahdat · Evgeny Andriyash · William Macready

Boltzmann machines are powerful distributions that have been shown to be an effective prior over binary latent variables in variational autoencoders (VAEs). However, previous methods for training discrete VAEs have used the evidence lower bound and not the tighter importance-weighted bound. We propose two approaches for relaxing Boltzmann machines to continuous distributions that permit training with importance-weighted bounds. These relaxations are based on generalized overlapping transformations and the Gaussian integral trick. Experiments on the MNIST and OMNIGLOT datasets show that these relaxations outperform previous discrete VAEs with Boltzmann priors. An implementation which reproduces these results is available.

Amortized Inference Regularization

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #7

Rui Shu · Hung Bui · Shengjia Zhao · Mykel J Kochenderfer · Stefano Ermon

The variational autoencoder (VAE) is a popular model for density estimation and representation learning. Canonically, the variational principle suggests to prefer an expressive inference model so that the variational approximation is accurate. However, it is often overlooked that an overly-expressive inference model can be detrimental to the test set performance of both the amortized posterior approximator and, more importantly, the generative density estimator. In this paper, we leverage the fact that VAEs rely on amortized inference and propose techniques for amortized inference regularization (AIR) that control the smoothness of the inference model. We demonstrate that, by applying AIR, it is possible to improve VAE generalization on both inference and generative performance. Our paper challenges the belief that amortized inference is simply a mechanism for approximating maximum likelihood training and illustrates that regularization of the amortization family provides a new direction for understanding and improving generalization in VAEs.

GumBolt: Extending Gumbel trick to Boltzmann priors

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #8

Amir H Khoshaman · Mohammad Amin

Boltzmann machines (BMs) are appealing candidates for powerful priors in variational autoencoders (VAEs), as they are capable of capturing nontrivial and multi-modal distributions over discrete variables. However, non-differentiability of the discrete units prohibits using the reparameterization trick, essential for low-noise back propagation. The Gumbel trick resolves this problem in a consistent way by relaxing the variables and distributions, but it is incompatible with BM priors. Here, we propose the GumBolt, a model that extends the Gumbel trick to BM priors in VAEs. GumBolt is significantly simpler than the recently proposed methods with BM prior and outperforms them by a considerable margin. It achieves state-of-the-art performance on permutation invariant MNIST and OMNIGLOT datasets in the scope of models with only discrete latent variables. Moreover, the performance can be further improved by allowing multi-sampled (importance-weighted) estimation of log-likelihood in training, which was not possible with previous models.

Constrained Generation of Semantically Valid Graphs via

Regularizing Variational Autoencoders

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #9

Tengfei Ma · Jie Chen · Cao Xiao

Deep generative models have achieved remarkable success in various data domains, including images, time series, and natural languages. There remain, however, substantial challenges for combinatorial structures, including graphs. One of the key challenges lies in the difficulty of ensuring semantic validity in context. For example, in molecular graphs, the number of bonding-electron pairs must not exceed the valence of an atom; whereas in protein interaction networks, two proteins may be connected only when they belong to the same or correlated gene ontology terms. These constraints are not easy to be incorporated into a generative model. In this work, we propose a regularization framework for variational autoencoders as a step toward semantic validity. We focus on the matrix representation of graphs and formulate penalty terms that regularize the output distribution of the decoder to encourage the satisfaction of validity constraints. Experimental results confirm a much higher likelihood of sampling valid graphs in our approach, compared with others reported in the literature.

Invertibility of Convolutional Generative Networks from Partial Measurements

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #10

Fangchang Ma · Ulas Ayaz · Sertac Karaman

In this work, we present new theoretical results on convolutional generative neural networks, in particular their invertibility (i.e., the recovery of input latent code given the network output). The study of network inversion problem is motivated by image inpainting and the mode collapse problem in training GAN. Network inversion is highly non-convex, and thus is typically computationally intractable and without optimality guarantees. However, we rigorously prove that, under some mild technical assumptions, the input of a two-layer convolutional generative network can be deduced from the network output efficiently using simple gradient descent. This new theoretical finding implies that the mapping from the low- dimensional latent space to the high-dimensional image space is bijective (i.e., one-to-one). In addition, the same conclusion holds even when the network output is only partially observed (i.e., with missing pixels). Our theorems hold for 2-layer convolutional generative network with ReLU as the activation function, but we demonstrate empirically that the same conclusion extends to multi-layer networks and networks with other activation functions, including the leaky ReLU, sigmoid and tanh.

Glow: Generative Flow with Invertible 1x1 Convolutions

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #11

Durk Kingma · Prafulla Dhariwal

Flow-based generative models are conceptually attractive due to tractability of the exact log-likelihood, tractability of exact latent-variable inference, and parallelizability of both training and synthesis. In this paper we propose Glow, a simple type of generative flow using invertible 1x1 convolution. Using our method we demonstrate a significant improvement in log-likelihood and qualitative sample quality. Perhaps most strikingly, we demonstrate that a generative model optimized towards the plain log-likelihood objective is capable of efficient synthesis of large and subjectively realistic-looking images.

Multimodal Generative Models for Scalable Weakly-Supervised Learning

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #12

Mike Wu · Noah Goodman

Multiple modalities often co-occur when describing natural phenomena. Learning a joint representation of these modalities should yield deeper and more useful representations.Previous generative approaches to multi-modal input either do not learn a joint distribution or require additional computation to handle missing data. Here, we introduce a multimodal variational autoencoder (MVAE) that uses a product-of-experts inference network and a sub-sampled training paradigm to solve the multi-modal inference problem. Notably, our model shares parameters to efficiently learn under any combination of missing modalities. We apply the MVAE on four datasets and match state-of-the-art performance using many fewer parameters. In addition, we show that the MVAE is directly applicable to weakly-supervised learning, and is robust to incomplete supervision. We then consider two case studies, one of learning image transformations---edge detection, colorization, segmentation---as a set of modalities, followed by one of machine translation between two languages. We find appealing results across this range of tasks.

IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #13

Huaibo Huang · zhihang li · Ran He · Zhenan Sun · Tieniu Tan

We present a novel introspective variational autoencoder (IntroVAE) model for synthesizing high-resolution photographic images. IntroVAE is capable of self-evaluating the quality of its generated samples and improving itself accordingly. Its inference and generator models are jointly trained in an introspective way. On one hand, the generator is required to reconstruct the input images from the noisy outputs of the inference model as normal VAEs. On the other hand, the inference model is encouraged to classify between the generated and real samples while the generator tries to fool it as GANs. These two famous generative frameworks are integrated in a simple yet efficient single-stream architecture that can be trained in a single stage. IntroVAE preserves the advantages of VAEs, such as stable training and nice latent manifold. Unlike most other hybrid models of VAEs and GANs, IntroVAE requires no extra discriminators, because the inference model itself serves as a discriminator to distinguish between the generated and real samples. Experiments demonstrate that our method produces high-resolution photo-realistic images (e.g., CELEBA images at (1024^{2})), which are comparable to or better than the state-of-the-art GANs.

Towards Text Generation with Adversarially Learned Neural Outlines

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #14

Sandeep Subramanian · Sai Rajeswar Mudumba · Alessandro Sordoni · Adam Trischler · Aaron Courville · Chris Pal

Recent progress in deep generative models has been fueled by two paradigms -- autoregressive and adversarial models. We propose a combination of both approaches with the goal of learning generative models of text. Our method first produces a high-level sentence outline and then generates words sequentially, conditioning on both the outline and the previous outputs. We generate outlines with an adversarial model trained to approximate the distribution of sentences in a latent space induced by general-purpose sentence encoders. This provides strong, informative conditioning for the autoregressive stage. Our quantitative evaluations suggests that conditioning information from generated outlines is able to guide the autoregressive model to produce realistic samples, comparable to maximum-likelihood trained language models, even at high temperatures with multinomial sampling. Qualitative results also demonstrate that this generative procedure yields natural-looking sentences and interpolations.

Unsupervised Image-to-Image Translation Using Domain-Specific Variational Information Bound

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #15

Hadi Kazemi · Sobhan Soleymani · Fariborz Taherkhani · Seyed Iranmanesh · Nasser Nasrabadi

Unsupervised image-to-image translation is a class of computer vision problems which aims at modeling conditional distribution of images in the target domain, given a set of unpaired images in the source and target domains. An image in the source domain might have multiple representations in the target domain. Therefore, ambiguity in modeling of the conditional distribution arises, specially when the images in the source and target domains come from different modalities. Current approaches mostly rely on simplifying assumptions to map both domains into a shared-latent space. Consequently, they are only able to model the domain-invariant information between the two modalities. These approaches cannot model domain-specific information which has no representation in the target domain. In this work, we propose an unsupervised image-to-image translation framework which maximizes a domain-specific variational information bound and learns the target domain-invariant representation of the two domain. The proposed framework makes it possible to map a single source image into multiple images in the target domain, utilizing several target domain-specific codes sampled randomly from the prior distribution, or extracted from reference images.

Adversarial Scene Editing: Automatic Object Removal from Weak Supervision

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #16

Rakshith R Shetty · Mario Fritz · Bernt Schiele

While great progress has been made recently in automatic image manipulation, it has been limited to object centric images like faces or structured scene datasets. In this work, we take a step towards general scene-level image editing by developing an automatic interaction-free object removal model. Our model learns to find and remove objects from general scene images using image-level labels and unpaired data in a generative adversarial network (GAN) framework. We achieve this with two key contributions: a two-stage editor architecture consisting of a mask generator and image in-painter that co-operate to remove objects, and a novel GAN based prior for the mask generator that allows us to flexibly incorporate knowledge about object shapes. We experimentally show on two datasets that our method effectively removes a wide variety of objects using weak supervision only.

Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #17

Boris Muzellec · Marco Cuturi

Embedding complex objects as vectors in low dimensional spaces is a longstanding problem in machine learning. We propose in this work an extension of that approach, which consists in embedding objects as elliptical probability distributions, namely distributions whose densities have elliptical level sets. We endow these measures with the 2-Wasserstein metric, with two important benefits: (i) For such measures, the squared 2-Wasserstein metric has a closed form, equal to a weighted sum of the squared Euclidean distance between means and the squared Bures metric between covariance matrices. The latter is a Riemannian metric between positive semi-definite matrices, which turns out to be Euclidean on a suitable factor representation of such matrices, which is valid on the entire geodesic between these matrices. (ii) The 2-Wasserstein distance boils down to the usual Euclidean metric when comparing Diracs, and therefore provides a natural framework to extend point embeddings. We show that for these reasons Wasserstein elliptical embeddings are more intuitive and yield tools that are better behaved numerically than the alternative choice of Gaussian embeddings with the Kullback-Leibler divergence. In particular, and unlike previous work based on the KL geometry, we learn elliptical distributions that are not necessarily diagonal. We demonstrate the advantages of elliptical embeddings by using them for visualization, to compute embeddings of words, and to reflect entailment or hypernymy.

Banach Wasserstein GAN

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #18

Jonas Adler · Sebastian Lunz

Wasserstein Generative Adversarial Networks (WGANs) can be used to generate realistic samples from complicated image distributions. The Wasserstein metric used in WGANs is based on a notion of distance between individual images, which induces a notion of distance between probability distributions of images. So far the community has considered ℓ^2 as the underlying distance. We generalize the theory of WGAN with gradient penalty to Banach spaces, allowing practitioners to select the features to emphasize in the generator. We further discuss the effect of some particular choices of underlying norms, focusing on Sobolev norms. Finally, we demonstrate a boost in performance for an appropriate choice of norm on CIFAR-10 and CelebA.

A Convex Duality Framework for GANs

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #19

Farzan Farnia · David Tse

Generative adversarial network (GAN) is a minimax game between a generator mimicking the true model and a discriminator distinguishing the samples produced by the generator from the real training samples. Given an unconstrained discriminator able to approximate any function, this game reduces to finding the generative model minimizing a divergence measure, e.g. the Jensen-Shannon (JS) divergence, to the data distribution. However, in practice the discriminator is constrained to be in a smaller class F such as neural nets. Then, a natural question is how the divergence minimization interpretation changes as we constrain F. In this work, we address this question by developing a convex duality framework for analyzing GANs. For a convex set F, this duality framework interprets the original GAN formulation as finding the generative model with minimum JS-divergence to the distributions penalized to match the moments of the data distribution, with the moments specified by the discriminators in F. We show that this interpretation more generally holds for f-GAN and Wasserstein GAN. As a byproduct, we apply the duality framework to a hybrid of f-divergence and Wasserstein distance. Unlike the f-divergence, we prove that the proposed hybrid divergence changes continuously with the generative model, which suggests regularizing the discriminator's Lipschitz constant in f-GAN and vanilla GAN. We numerically evaluate the power of the suggested regularization schemes for improving GAN's training performance.

On the Convergence and Robustness of Training GANs with Regularized Optimal Transport

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #20

Maziar Sanjabi · Jimmy Ba · Meisam Razaviyayn · Jason Lee

Generative Adversarial Networks (GANs) are one of the most practical methods for learning data distributions. A popular GAN formulation is based on the use of Wasserstein distance as a metric between probability distributions. Unfortunately, minimizing the Wasserstein distance between the data distribution and the generative model distribution is a computationally challenging problem as its objective is non-convex, non-smooth, and even hard to compute. In this work, we show that obtaining gradient information of the smoothed Wasserstein GAN formulation, which is based on regularized Optimal Transport (OT), is computationally effortless and hence one can apply first order optimization methods to minimize this objective. Consequently, we establish theoretical convergence guarantee to stationarity for a proposed class of GAN optimization algorithms. Unlike the original non-smooth formulation, our algorithm only requires solving the discriminator to approximate optimality. We apply our method to learning MNIST digits as well as CIFAR-10 images. Our experiments show that our method is computationally efficient and generates images comparable to the state of the art algorithms given the same architecture and computational power.

On gradient regularizers for MMD GANs

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #21

Michael Arbel · Dougal Sutherland · Mikołaj Bińkowski · Arthur Gretton

We propose a principled method for gradient-based regularization of the critic of GAN-like models trained by adversarially optimizing the kernel of a Maximum Mean Discrepancy (MMD). We show that controlling the gradient of the critic is vital to having a sensible loss function, and devise a method to enforce exact, analytical gradient constraints at no additional cost compared to existing approximate techniques based on additive regularizers. The new loss function is provably continuous, and experiments show that it stabilizes and accelerates training, giving image generation models that outperform state-of-the art methods on 160×160 CelebA and 64×64 unconditional ImageNet.

PacGAN: The power of two samples in generative adversarial networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #22

Zinan Lin · Ashish Khetan · Giulia Fanti · Sewoong Oh

Generative adversarial networks (GANs) are a technique for learning generative models of complex data distributions from samples. Despite remarkable advances in generating realistic images, a major shortcoming of GANs is the fact that they tend to produce samples with little diversity, even when trained on diverse datasets. This phenomenon, known as mode collapse, has been the focus of much recent work. We study a principled approach to handling mode collapse, which we call packing. The main idea is to modify the discriminator to make decisions based on multiple samples from the same class, either real or artificially generated. We draw analysis tools from binary hypothesis testing---in particular the seminal result of Blackwell---to prove a fundamental connection between packing and mode collapse. We show that packing naturally penalizes generators with mode collapse, thereby favoring generator distributions with less mode collapse during the training process. Numerical experiments on benchmark datasets suggest that packing provides significant improvements.

Are GANs Created Equal? A Large-Scale Study

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #23

Mario Lucic · Karol Kurach · Marcin Michalski · Sylvain Gelly · Olivier Bousquet

Generative adversarial networks (GAN) are a powerful subclass of generative models. Despite a very rich research activity leading to numerous interesting GAN algorithms, it is still very hard to assess which algorithm(s) perform better than others. We conduct a neutral, multi-faceted large-scale empirical study on state-of-the art models and evaluation measures. We find that most models can reach similar scores with enough hyperparameter optimization and random restarts. This suggests that improvements can arise from a higher computational budget and tuning more than fundamental algorithmic changes. To overcome some limitations of the current metrics, we also propose several data sets on which precision and recall can be computed. Our experimental results suggest that future GAN research should be based on more systematic and objective evaluation procedures. Finally, we did not find evidence that any of the tested algorithms consistently outperforms the non-saturating GAN introduced in \cite{goodfellow2014generative}.

Disconnected Manifold Learning for Generative Adversarial Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #24

Mahyar Khayatkhoei · Maneesh K. Singh · Ahmed Elgammal

Natural images may lie on a union of disjoint manifolds rather than one globally connected manifold, and this can cause several difficulties for the training of common Generative Adversarial Networks (GANs). In this work, we first show that single generator GANs are unable to correctly model a distribution supported on a disconnected manifold, and investigate how sample quality, mode dropping and local convergence are affected by this. Next, we show how using a collection of generators can address this problem, providing new insights into the success of such multi-generator GANs. Finally, we explain the serious issues caused by considering a fixed prior over the collection of generators and propose a novel approach for learning the prior and inferring the necessary number of generators without any supervision. Our proposed modifications can be applied on top of any other GAN model to enable learning of distributions supported on disconnected manifolds. We conduct several experiments to illustrate the aforementioned shortcoming of GANs, its consequences in practice, and the effectiveness of our proposed modifications in alleviating these issues.

Hessian-based Analysis of Large Batch Training and Robustness to Adversaries

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #25

Zhewei Yao · Amir Gholami · Qi Lei · Kurt Keutzer · Michael W Mahoney

Large batch size training of Neural Networks has been shown to incur accuracy loss when trained with the current methods. The exact underlying reasons for this are still not completely understood. Here, we study large batch size training through the lens of the Hessian operator and robust optimization. In particular, we perform a Hessian based study to analyze exactly how the landscape of the loss function changes when training with large batch size. We compute the true Hessian spectrum, without approximation, by back-propagating the second derivative. Extensive experiments on multiple networks show that saddle-points are not the cause for generalization gap of large batch size training, and the results consistently show that large batch converges to points with noticeably higher Hessian spectrum. Furthermore, we show that robust training allows one to favor flat areas, as points with large Hessian spectrum show poor robustness to adversarial perturbation. We further study this relationship, and provide empirical and theoretical proof that the inner loop for robust training is a saddle-free optimization problem \textit{almost everywhere}. We present detailed experiments with five different network architectures, including a residual network, tested on MNIST, CIFAR-10/100 datasets.

Fast and Effective Robustness Certification

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #26

Gagandeep Singh · Timon Gehr · Matthew Mirman · Markus Püschel · Martin Vechev

We present a new method and system, called DeepZ, for certifying neural network robustness based on abstract interpretation. Compared to state-of-the-art automated verifiers for neural networks, DeepZ: (i) handles ReLU, Tanh and Sigmoid activation functions, (ii) supports feedforward and convolutional architectures, (iii) is significantly more scalable and precise, and (iv) and is sound with respect to floating point arithmetic. These benefits are due to carefully designed approximations tailored to the setting of neural networks. As an example, DeepZ achieves a verification accuracy of 97% on a large network with 88,500 hidden units under L_{∞} attack with $\epsilon = 0.1$ with an average runtime of 133 seconds.

Graphical Generative Adversarial Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #27

Chongxuan LI · Max Welling · Jun Zhu · Bo Zhang

We propose Graphical Generative Adversarial Networks (Graphical-GAN) to model structured data. Graphical-GAN conjoins the power of Bayesian networks on compactly representing the dependency structures among random variables and that of generative adversarial networks on learning expressive dependency functions. We introduce a structured recognition model to infer the posterior distribution of latent variables given observations. We generalize the Expectation Propagation (EP) algorithm to learn the generative model and recognition model jointly. Finally, we present two important instances of Graphical-GAN, i.e. Gaussian Mixture GAN (GMGAN) and State Space GAN (SSGAN), which can successfully learn the discrete and temporal structures on visual datasets, respectively.

Deep Defense: Training DNNs with Improved Adversarial Robustness

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #28

Ziang Yan · Yiwen Guo · Changshui Zhang

Despite the efficacy on a variety of computer vision tasks, deep neural networks (DNNs) are vulnerable to adversarial attacks, limiting their applications in security-critical systems. Recent works have shown the possibility of generating imperceptibly perturbed image inputs (a.k.a., adversarial examples) to fool well-trained DNN classifiers into making arbitrary predictions. To address this problem, we propose a training recipe named "deep defense". Our core idea is to integrate an adversarial perturbation-based regularizer into the classification objective, such that the obtained models learn to resist potential attacks, directly and precisely. The whole optimization problem is solved just like training a recursive network. Experimental results demonstrate that our method outperforms training with adversarial/Parseval regularizations by large margins on various datasets (including MNIST, CIFAR-10 and ImageNet) and different DNN architectures. Code and models for reproducing our results are available at https://github.com/ZiangYan/deepdefense.pytorch.

Learning to Repair Software Vulnerabilities with Generative Adversarial Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #29

Jacob Harer · Onur Ozdemir · Tomo Lazovich · Christopher Reale · Rebecca Russell · Louis Kim · peter chin

Motivated by the problem of automated repair of software vulnerabilities, we propose an adversarial learning approach that maps from one discrete source domain to another target domain without requiring paired labeled examples or source and target domains to be bijections. We demonstrate that the proposed adversarial learning approach is an effective technique for repairing software vulnerabilities, performing close to seq2seq approaches that require labeled pairs. The proposed Generative Adversarial Network approach is application-agnostic in that it can be applied to other problems similar to code repair, such as grammar correction or sentiment translation.

Memory Replay GANs: Learning to Generate New Categories without Forgetting

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #30

Chenshen Wu · Luis Herranz · Xialei Liu · yaxing wang · Joost van de Weijer · Bogdan Raducanu

Previous works on sequential learning address the problem of forgetting in discriminative models. In this paper we consider the case of generative models. In particular, we investigate generative adversarial networks (GANs) in the task of learning new categories in a sequential fashion. We first show that sequential fine tuning renders the network unable to properly generate images from previous categories (i.e. forgetting). Addressing this problem, we propose Memory Replay GANs (MeRGANs), a conditional GAN framework that integrates a memory replay generator. We study two methods to prevent forgetting by leveraging these replays, namely joint training with replay and replay alignment. Qualitative and quantitative experimental results in MNIST, SVHN and LSUN datasets show that our memory replay approach can generate competitive images while significantly mitigating the forgetting of previous categories.

Unsupervised Attention-guided Image-to-Image Translation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #31

Youssef Alami Mejjati · Christian Richardt · James Tompkin · Darren Cosker · Kwang In Kim

Current unsupervised image-to-image translation techniques struggle to focus their attention on individual objects without altering the background or the way multiple objects interact within a scene. Motivated by the important role of attention in human perception, we tackle this limitation by introducing unsupervised attention mechanisms which are jointly adversarially trained with the generators and discriminators. We empirically demonstrate that our approach is able to attend to relevant regions in the image without requiring any additional supervision, and that by doing so it achieves more realistic mappings compared to recent approaches.

Conditional Adversarial Domain Adaptation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #32

Mingsheng Long · ZHANGJIE CAO · Jianmin Wang · Michael Jordan

Adversarial learning has been embedded into deep networks to learn disentangled and transferable representations for domain adaptation. Existing adversarial domain adaptation methods may struggle to align different domains of multimodal distributions that are native in classification problems. In this paper, we present conditional adversarial domain adaptation, a principled framework that conditions the adversarial adaptation models on discriminative information conveyed in the classifier predictions. Conditional domain adversarial networks (CDANs) are designed with two novel conditioning strategies: multilinear conditioning that captures the cross-covariance between feature representations and classifier predictions to improve the discriminability, and entropy conditioning that controls the uncertainty of classifier predictions to guarantee the transferability. Experiments testify that the proposed approach exceeds the state-of-the-art results on five benchmark datasets.

Video-to-Video Synthesis

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #33

Ting-Chun Wang · Ming-Yu Liu · Jun-Yan Zhu · Nikolai Yakovenko · Andrew Tao · Jan Kautz · Bryan Catanzaro

We study the problem of video-to-video synthesis, whose goal is to learn a mapping function from an input source video (e.g., a sequence of semantic segmentation masks) to an output photorealistic video that precisely depicts the content of the source video. While its image counterpart, the image-to-image translation problem, is a popular topic, the video-to-video synthesis problem is less explored in the literature. Without modeling temporal dynamics, directly applying existing image synthesis approaches to an input video often results in temporally incoherent videos of low visual quality. In this paper, we propose a video-to-video synthesis approach under the generative adversarial learning framework. Through carefully-designed generators and discriminators, coupled with a spatio-temporal adversarial objective, we achieve high-resolution, photorealistic, temporally coherent video results on a diverse set of input formats including segmentation masks, sketches, and poses. Experiments on multiple benchmarks show the advantage of our method compared to strong baselines. In particular, our model is capable of synthesizing 2K resolution videos of street scenes up to 30 seconds long, which significantly advances the state-of-the-art of video synthesis. Finally, we apply our method to future video prediction, outperforming several competing systems. Code, models, and more results are available at our website: https://github.com/NVIDIA/vid2vid. (Please use Adobe Reader to see the embedded videos in the paper.)

Generalized Zero-Shot Learning with Deep Calibration Network

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #34

Shichen Liu · Mingsheng Long · Jianmin Wang · Michael Jordan

A technical challenge of deep learning is recognizing target classes without seen data. Zero-shot learning leverages semantic representations such as attributes or class prototypes to bridge source and target classes. Existing standard zero-shot learning methods may be prone to overfitting the seen data of source classes as they are blind to the semantic representations of target classes. In this paper, we study generalized zero-shot learning that assumes accessible to target classes for unseen data during training, and prediction on unseen data is made by searching on both source and target classes. We propose a novel Deep Calibration Network (DCN) approach towards this generalized zero-shot learning paradigm, which enables simultaneous calibration of deep networks on the confidence of source classes and uncertainty of target classes. Our approach maps visual features of images and semantic representations of class prototypes to a common embedding space such that the compatibility of seen data to both source and target classes are maximized. We show superior accuracy of our approach over the state of the art on benchmark datasets for generalized zero-shot learning, including AwA, CUB, SUN, and aPY.

Low-shot Learning via Covariance-Preserving Adversarial Augmentation Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #35

Hang Gao · Zheng Shou · Alireza Zareian · Hanwang Zhang · Shih-Fu Chang

Deep neural networks suffer from over-fitting and catastrophic forgetting when trained with small data. One natural remedy for this problem is data augmentation, which has been recently shown to be effective. However, previous works either assume that intra-class variances can always be generalized to new classes, or employ naive generation methods to hallucinate finite examples without modeling their latent distributions. In this work, we propose Covariance-Preserving Adversarial Augmentation Networks to overcome existing limits of low-shot learning. Specifically, a novel Generative Adversarial Network is designed to model the latent distribution of each novel class given its related base counterparts. Since direct estimation on novel classes can be inductively biased, we explicitly preserve covariance information as the ``variability'' of base examples during the generation process. Empirical results show that our model can generate realistic yet diverse examples, leading to substantial improvements on the ImageNet benchmark over the state of the art.

Trading robust representations for sample complexity through self-supervised visual experience

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #36

Andrea Tacchetti · Stephen Voinea · Georgios Evangelopoulos

Learning in small sample regimes is among the most remarkable features of the human perceptual system. This ability is related to robustness to transformations, which is acquired through visual experience in the form of weak- or self-supervision during development. We explore the idea of allowing artificial systems to learn representations of visual stimuli through weak supervision prior to downstream supervised tasks. We introduce a novel loss function for representation learning using unlabeled image sets and video sequences, and experimentally demonstrate that these representations support one-shot learning and reduce the sample complexity of multiple recognition tasks. We establish the existence of a trade-off between the sizes of weakly supervised, automatically obtained from video sequences, and fully supervised data sets. Our results suggest that equivalence sets other than class labels, which are abundant in unlabeled visual experience, can be used for self-supervised learning of semantically relevant image embeddings.

TADAM: Task dependent adaptive metric for improved few-shot learning

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #37

Boris Oreshkin · Pau Rodríguez López · Alexandre Lacoste

Few-shot learning has become essential for producing models that generalize from few examples. In this work, we identify that metric scaling and metric task conditioning are important to improve the performance of few-shot algorithms. Our analysis reveals that simple metric scaling completely changes the nature of few-shot algorithm parameter updates. Metric scaling provides improvements up to 14% in accuracy for certain metrics on the mini-Imagenet 5-way 5-shot classification task. We further propose a simple and effective way of conditioning a learner on the task sample set, resulting in learning a task-dependent metric space. Moreover, we propose and empirically test a practical end-to-end optimization procedure based on auxiliary task co-training to learn a task-dependent metric space. The resulting few-shot learning model based on the task-dependent scaled metric achieves state of the art on mini-Imagenet. We confirm these results on another few-shot dataset that we introduce in this paper based on CIFAR100.

FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #38

Shuyang Sun · Jiangmiao Pang · Jianping Shi · Shuai Yi · Wanli Ouyang

The basic principles in designing convolutional neural network (CNN) structures for predicting objects on different levels, e.g., image-level, region-level, and pixel-level, are diverging. Generally, network structures designed specifically for image classification are directly used as default backbone structure for other tasks including detection and segmentation, but there is seldom backbone structure designed under the consideration of unifying the advantages of networks designed for pixel-level or region-level predicting tasks, which may require very deep features with high resolution. Towards this goal, we design a fish-like network, called FishNet. In FishNet, the information of all resolutions is preserved and refined for the final task. Besides, we observe that existing works still cannot \emph{directly} propagate the gradient information from deep layers to shallow layers. Our design can better handle this problem. Extensive experiments have been conducted to demonstrate the remarkable performance of the FishNet. In particular, on ImageNet-1k, the accuracy of FishNet is able to surpass the performance of DenseNet and ResNet with fewer parameters. FishNet was applied as one of the modules in the winning entry of the COCO Detection 2018 challenge. The code is available at https://github.com/kevin-ssy/FishNet.

Batch-Instance Normalization for Adaptively Style-Invariant Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #39

Hyeonseob Nam · Hyo-Eun Kim

Real-world image recognition is often challenged by the variability of visual styles including object textures, lighting conditions, filter effects, etc. Although these variations have been deemed to be implicitly handled by more training data and deeper networks, recent advances in image style transfer suggest that it is also possible to explicitly manipulate the style information. Extending this idea to general visual recognition problems, we present Batch-Instance Normalization (BIN) to explicitly normalize unnecessary styles from images. Considering certain style features play an essential role in discriminative tasks, BIN learns to selectively normalize only disturbing styles while preserving useful styles. The proposed normalization module is easily incorporated into existing network architectures such as Residual Networks, and surprisingly improves the recognition performance in various scenarios. Furthermore, experiments verify that BIN effectively adapts to completely different tasks like object classification and style transfer, by controlling the trade-off between preserving and removing style variations. BIN can be implemented with only a few lines of code using popular deep learning frameworks.

A^2-Nets: Double Attention Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #40

Yunpeng Chen · Yannis Kalantidis · Jianshu Li · Shuicheng Yan · Jiashi Feng

Learning to capture long-range relations is fundamental to image/video recognition. Existing CNN models generally rely on increasing depth to model such relations which is highly inefficient. In this work, we propose the “double attention block”, a novel component that aggregates and propagates informative global features from the entire spatio-temporal space of input images/videos, enabling subsequent convolution layers to access features from the entire space efficiently. The component is designed with a double attention mechanism in two steps, where the first step gathers features from the entire space into a compact set through second-order attention pooling and the second step adaptively selects and distributes features to each location via another attention. The proposed double attention block is easy to adopt and can be plugged into existing deep neural networks conveniently. We conduct extensive ablation studies and experiments on both image and video recognition tasks for evaluating its performance. On the image recognition task, a ResNet-50 equipped with our double attention blocks outperforms a much larger ResNet-152 architecture on ImageNet-1k dataset with over 40% less the number of parameters and less FLOPs. On the action recognition task, our proposed model achieves the state-of-the-art results on the Kinetics and UCF-101 datasets with significantly higher efficiency than recent works.

Pelee: A Real-Time Object Detection System on Mobile Devices

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #41

Jun Wang · Tanner Bohn · Charles Ling

An increasing need of running Convolutional Neural Network (CNN) models on mobile devices with limited computing power and memory resource encourages studies on efficient model design. A number of efficient architectures have been proposed in recent years, for example, MobileNet, ShuffleNet, and MobileNetV2. However, all these models are heavily dependent on depthwise separable convolution which lacks efficient implementation in most deep learning frameworks. In this study, we propose an efficient architecture named PeleeNet, which is built with conventional convolution instead. On ImageNet ILSVRC 2012 dataset, our proposed PeleeNet achieves a higher accuracy and 1.8 times faster speed than MobileNet and MobileNetV2 on NVIDIA TX2. Meanwhile, PeleeNet is only 66% of the model size of MobileNet. We then propose a real-time object detection system by combining PeleeNet with Single Shot MultiBox Detector (SSD) method and optimizing the architecture for fast speed. Our proposed detection system, named Pelee, achieves 76.4% mAP (mean average precision) on PASCAL VOC2007 and 22.4 mAP on MS COCO dataset at the speed of 23.6 FPS on iPhone 8 and 125 FPS on NVIDIA TX2. The result on COCO outperforms YOLOv2 in consideration of a higher precision, 13.6 times lower computational cost and 11.3 times smaller model size. The code and models are open sourced.

PointCNN: Convolution On X-Transformed Points

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #42

Yangyan Li · Rui Bu · Mingchao Sun · Wei Wu · Xinhan Di · Baoquan Chen

We present a simple and general framework for feature learning from point cloud. The key to the success of CNNs is the convolution operator that is capable of leveraging spatially-local correlation in data represented densely in grids (e.g. images). However, point cloud are irregular and unordered, thus a direct convolving of kernels against the features associated with the points will result in deserting the shape information while being variant to the orders. To address these problems, we propose to learn a X-transformation from the input points, which is used for simultaneously weighting the input features associated with the points and permuting them into latent potentially canonical order. Then element-wise product and sum operations of typical convolution operator are applied on the X-transformed features. The proposed method is a generalization of typical CNNs into learning features from point cloud, thus we call it PointCNN. Experiments show that PointCNN achieves on par or better performance than state-of-the-art methods on multiple challenging benchmark datasets and tasks.

Deep Neural Networks with Box Convolutions

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #43

Egor Burkov · Victor Lempitsky

Box filters computed using integral images have been part of the computer vision toolset for a long time. Here, we show that a convolutional layer that computes box filter responses in a sliding manner can be used within deep architectures, whereas the dimensions and the offsets of the sliding boxes in such a layer can be learned as part of an end-to-end loss minimization. Crucially, the training process can make the size of the boxes in such a layer arbitrarily large without incurring extra computational cost and without the need to increase the number of learnable parameters. Due to its ability to integrate information over large boxes, the new layer facilitates long-range propagation of information and leads to the efficient increase of the receptive fields of downstream units in the network. By incorporating the new layer into existing architectures for semantic segmentation, we are able to achieve both the increase in segmentation accuracy as well as the decrease in the computational cost and the number of learnable parameters.

An intriguing failing of convolutional neural networks and the CoordConv solution

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #44

Rosanne Liu · Joel Lehman · Piero Molino · Felipe Petroski Such · Eric Frank · Alex Sergeev · Jason Yosinski

Few ideas have enjoyed as large an impact on deep learning as convolution. For any problem involving pixels or spatial representations, common intuition holds that convolutional neural networks may be appropriate. In this paper we show a striking counterexample to this intuition via the seemingly trivial coordinate transform problem, which simply requires learning a mapping between coordinates in (x,y) Cartesian space and coordinates in one-hot pixel space. Although convolutional networks would seem appropriate for this task, we show that they fail spectacularly. We demonstrate and carefully analyze the failure first on a toy problem, at which point a simple fix becomes obvious. We call this solution CoordConv, which works by giving convolution access to its own input coordinates through the use of extra coordinate channels. Without sacrificing the computational and parametric efficiency of ordinary convolution, CoordConv allows networks to learn either complete translation invariance or varying degrees of translation dependence, as required by the end task. CoordConv solves the coordinate transform problem with perfect generalization and 150 times faster with 10--100 times fewer parameters than convolution. This stark contrast raises the question: to what extent has this inability of convolution persisted insidiously inside other tasks, subtly hampering performance from within? A complete answer to this question will require further investigation, but we show preliminary evidence that swapping convolution for CoordConv can improve models on a diverse set of tasks. Using CoordConv in a GAN produced less mode collapse as the transform between high-level spatial latents and pixels becomes easier to learn. A Faster R-CNN detection model trained on MNIST detection showed 24% better IOU when using CoordConv, and in the Reinforcement Learning (RL) domain agents playing Atari games benefit significantly from the use of CoordConv layers.

3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #45

Maurice Weiler · Wouter Boomsma · Mario Geiger · Max Welling · Taco Cohen

We present a convolutional network that is equivariant to rigid body motions. The model uses scalar-, vector-, and tensor fields over 3D Euclidean space to represent data, and equivariant convolutions to map between such representations. These SE(3)-equivariant convolutions utilize kernels which are parameterized as a linear combination of a complete steerable kernel basis, which is derived analytically in this paper. We prove that equivariant convolutions are the most general equivariant linear maps between fields over R^3. Our experimental results confirm the effectiveness of 3D Steerable CNNs for the problem of amino acid propensity prediction and protein structure classification, both of which have inherent SE(3) symmetry.

Moonshine: Distilling with Cheap Convolutions

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #46

Elliot J. Crowley · Gavin Gray · Amos Storkey

Many engineers wish to deploy modern neural networks in memory-limited settings; but the development of flexible methods for reducing memory use is in its infancy, and there is little knowledge of the resulting cost-benefit. We propose structural model distillation for memory reduction using a strategy that produces a student architecture that is a simple transformation of the teacher architecture: no redesign is needed, and the same hyperparameters can be used. Using attention transfer, we provide Pareto curves/tables for distillation of residual networks with four benchmark datasets, indicating the memory versus accuracy payoff. We show that substantial memory savings are possible with very little loss of accuracy, and confirm that distillation provides student network performance that is better than training that student architecture directly on data.

Kalman Normalization: Normalizing Internal Representations Across Network Layers

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #47

Guangrun Wang · jiefeng peng · Ping Luo · Xinjiang Wang · Liang Lin

As an indispensable component, Batch Normalization (BN) has successfully improved the training of deep neural networks (DNNs) with mini-batches, by normalizing the distribution of the internal representation for each hidden layer. However, the effectiveness of BN would diminish with the scenario of micro-batch (e.g. less than 4 samples in a mini-batch), since the estimated statistics in a mini-batch are not reliable with insufficient samples. This limits BN's room in training larger models on segmentation, detection, and video-related problems, which require small batches constrained by memory consumption. In this paper, we present a novel normalization method, called Kalman Normalization (KN), for improving and accelerating the training of DNNs, particularly under the context of micro-batches. Specifically, unlike the existing solutions treating each hidden layer as an isolated system, KN treats all the layers in a network as a whole system, and estimates the statistics of a certain layer by considering the distributions of all its preceding layers, mimicking the merits of Kalman Filtering. On ResNet50 trained in ImageNet, KN has 3.4% lower error than its BN counterpart when using a batch size of 4; Even when using typical batch sizes, KN still maintains an advantage over BN while other BN variants suffer a performance degradation. Moreover, KN can be naturally generalized to many existing normalization variants to obtain gains, e.g. equipping Group Normalization with Group Kalman Normalization (GKN). KN can outperform BN and its variants for large scale object detection and segmentation task in COCO 2017.

SplineNets: Continuous Neural Decision Graphs

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #48

Cem Keskin · Shahram Izadi

We present SplineNets, a practical and novel approach for using conditioning in convolutional neural networks (CNNs). SplineNets are continuous generalizations of neural decision graphs, and they can dramatically reduce runtime complexity and computation costs of CNNs, while maintaining or even increasing accuracy. Functions of SplineNets are both dynamic (i.e., conditioned on the input) and hierarchical (i.e.,conditioned on the computational path). SplineNets employ a unified loss function with a desired level of smoothness over both the network and decision parameters, while allowing for sparse activation of a subset of nodes for individual samples. In particular, we embed infinitely many function weights (e.g. filters) on smooth, low dimensional manifolds parameterized by compact B-splines, which are indexed by a position parameter. Instead of sampling from a categorical distribution to pick a branch, samples choose a continuous position to pick a function weight. We further show that by maximizing the mutual information between spline positions and class labels, the network can be optimally utilized and specialized for classification tasks. Experiments show that our approach can significantly increase the accuracy of ResNets with negligible cost in speed, matching the precision of a 110 level ResNet with a 32 level SplineNet.

CapProNet: Deep Feature Learning via Orthogonal Projections onto Capsule Subspaces

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #49

Liheng Zhang · Marzieh Edraki · Guo-Jun Qi

In this paper, we formalize the idea behind capsule nets of using a capsule vector rather than a neuron activation to predict the label of samples. To this end, we propose to learn a group of capsule subspaces onto which an input feature vector is projected. Then the lengths of resultant capsules are used to score the probability of belonging to different classes. We train such a Capsule Projection Network (CapProNet) by learning an orthogonal projection matrix for each capsule subspace, and show that each capsule subspace is updated until it contains input feature vectors corresponding to the associated class. With low dimensionality of capsule subspace as well as an iterative method to estimate the matrix inverse, only a small negligible computing overhead is incurred to train the network. Experiment results on image datasets show the presented network can greatly improve the performance of state-of-the-art Resnet backbones by $10-20\%$ with almost the same computing cost.

Which Neural Net Architectures Give Rise to Exploding and Vanishing Gradients?

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #50

Boris Hanin

We give a rigorous analysis of the statistical behavior of gradients in a randomly initialized fully connected network N with ReLU activations. Our results show that the empirical variance of the squares of the entries in the input-output Jacobian of N is exponential in a simple architecture-dependent constant beta, given by the sum of the reciprocals of the hidden layer widths. When beta is large, the gradients computed by N at initialization vary wildly. Our approach complements the mean field theory analysis of random networks. From this point of view, we rigorously compute finite width corrections to the statistics of gradients at the edge of chaos.

Exact natural gradient in deep linear networks and its application to the nonlinear case

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #51

Alberto Bernacchia · Mate Lengyel · Guillaume Hennequin

Stochastic gradient descent (SGD) remains the method of choice for deep learning, despite the limitations arising for ill-behaved objective functions. In cases where it could be estimated, the natural gradient has proven very effective at mitigating the catastrophic effects of pathological curvature in the objective function, but little is known theoretically about its convergence properties, and it has yet to find a practical implementation that would scale to very deep and large networks. Here, we derive an exact expression for the natural gradient in deep linear networks, which exhibit pathological curvature similar to the nonlinear case. We provide for the first time an analytical solution for its convergence rate, showing that the loss decreases exponentially to the global minimum in parameter space. Our expression for the natural gradient is surprisingly simple, computationally tractable, and explains why some approximations proposed previously work well in practice. This opens new avenues for approximating the natural gradient in the nonlinear case, and we show in preliminary experiments that our online natural gradient descent outperforms SGD on MNIST autoencoding while sharing its computational simplicity.

Fast Approximate Natural Gradient Descent in a Kronecker Factored Eigenbasis

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #52

Thomas George · César Laurent · Xavier Bouthillier · Nicolas Ballas · Pascal Vincent

Optimization algorithms that leverage gradient covariance information, such as variants of natural gradient descent (Amari, 1998), offer the prospect of yielding more effective descent directions. For models with many parameters, the covari- ance matrix they are based on becomes gigantic, making them inapplicable in their original form. This has motivated research into both simple diagonal approxima- tions and more sophisticated factored approximations such as KFAC (Heskes, 2000; Martens & Grosse, 2015; Grosse & Martens, 2016). In the present work we draw inspiration from both to propose a novel approximation that is provably better than KFAC and amendable to cheap partial updates. It consists in tracking a diagonal variance, not in parameter coordinates, but in a Kronecker-factored eigenbasis, in which the diagonal approximation is likely to be more effective.

Experiments show improvements over KFAC in optimization speed for several deep network architectures.

Post: Device Placement with Cross-Entropy Minimization and Proximal Policy Optimization

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #53

Yuanxiang Gao · Li Chen · Baochun Li

Training deep neural networks requires an exorbitant amount of computation resources, including a heterogeneous mix of GPU and CPU devices. It is critical to place operations in a neural network on these devices in an optimal way, so that the training process can complete within the shortest amount of time. The state-of-the-art uses reinforcement learning to learn placement skills by repeatedly performing Monte-Carlo experiments. However, due to its equal treatment of placement samples, we argue that there remains ample room for significant improvements. In this paper, we propose a new joint learning algorithm, called Post, that integrates cross-entropy minimization and proximal policy optimization to achieve theoretically guaranteed optimal efficiency. In order to incorporate the cross-entropy method as a sampling technique, we propose to represent placements using discrete probability distributions, which allows us to estimate an optimal probability mass by maximal likelihood estimation, a powerful tool with the best possible efficiency. We have implemented Post in the Google Cloud platform, and our extensive experiments with several popular neural network training benchmarks have demonstrated clear evidence of superior performance: with the same amount of learning time, it leads to placements that have training times up to 63.7% shorter over the state-of-the-art.

Paraphrasing Complex Network: Network Compression via Factor Transfer

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #54

Jangho Kim · Seonguk Park · Nojun Kwak

Many researchers have sought ways of model compression to reduce the size of a deep neural network (DNN) with minimal performance degradation in order to use DNNs in embedded systems. Among the model compression methods, a method called knowledge transfer is to train a student network with a stronger teacher network. In this paper, we propose a novel knowledge transfer method which uses convolutional operations to paraphrase teacher's knowledge and to translate it for the student. This is done by two convolutional modules, which are called a paraphraser and a translator. The paraphraser is trained in an unsupervised manner to extract the teacher factors which are defined as paraphrased information of the teacher network. The translator located at the student network extracts the student factors and helps to translate the teacher factors by mimicking them. We observed that our student network trained with the proposed factor transfer method outperforms the ones trained with conventional knowledge transfer methods.

Learning Compressed Transforms with Low Displacement Rank

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #55

Anna Thomas · Albert Gu · Tri Dao · Atri Rudra · Christopher Ré

The low displacement rank (LDR) framework for structured matrices represents a matrix through two displacement operators and a low-rank residual. Existing use of LDR matrices in deep learning has applied fixed displacement operators encoding forms of shift invariance akin to convolutions. We introduce a rich class of LDR matrices with more general displacement operators, and explicitly learn over both the operators and the low-rank component. This class generalizes several previous constructions while preserving compression and efficient computation. We prove bounds on the VC dimension of multi-layer neural networks with structured weight matrices and show empirically that our compact parameterization can reduce the sample complexity of learning. When replacing weight layers in fully-connected, convolutional, and recurrent neural networks for image classification and language modeling tasks, our new classes exceed the accuracy of existing compression approaches, and on some tasks even outperform general unstructured layers while using more than 20x fewer parameters.

Knowledge Distillation by On-the-Fly Native Ensemble

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #56

xu lan · Xiatian Zhu · Shaogang Gong

Knowledge distillation is effective to train the small and generalisable network models for meeting the low-memory and fast running requirements. Existing offline distillation methods rely on a strong pre-trained teacher, which enables favourable knowledge discovery and transfer but requires a complex two-phase training procedure. Online counterparts address this limitation at the price of lacking a high-capacity teacher. In this work, we present an On-the-fly Native Ensemble (ONE) learning strategy for one-stage online distillation. Specifically, ONE only trains a single multi-branch network while simultaneously establishing a strong teacher on-the-fly to enhance the learning of target network. Extensive evaluations show that ONE improves the generalisation performance of a variety of deep neural networks more significantly than alternative methods on four image classification dataset: CIFAR10, CIFAR100, SVHN, and ImageNet, whilst having the computational efficiency advantages.

Scalable methods for 8-bit training of neural networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #57

Ron Banner · Itay Hubara · Elad Hoffer · Daniel Soudry

Quantized Neural Networks (QNNs) are often used to improve network efficiency during the inference phase, i.e. after the network has been trained. Extensive research in the field suggests many different quantization schemes. Still, the number of bits required, as well as the best quantization scheme, are yet unknown. Our theoretical analysis suggests that most of the training process is robust to substantial precision reduction, and points to only a few specific operations that require higher precision. Armed with this knowledge, we quantize the model parameters, activations and layer gradients to 8-bit, leaving at higher precision only the final step in the computation of the weight gradients. Additionally, as QNNs require batch-normalization to be trained at high precision, we introduce Range Batch-Normalization (BN) which has significantly higher tolerance to quantization noise and improved computational complexity. Our simulations show that Range BN is equivalent to the traditional batch norm if a precise scale adjustment, which can be approximated analytically, is applied. To the best of the authors' knowledge, this work is the first to quantize the weights, activations, as well as a substantial volume of the gradients stream, in all layers (including batch normalization) to 8-bit while showing state-of-the-art results over the ImageNet-1K dataset.

Training Deep Models Faster with Robust, Approximate Importance Sampling

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #58

Tyler Johnson · Carlos Guestrin

In theory, importance sampling speeds up stochastic gradient algorithms for supervised learning by prioritizing training examples. In practice, the cost of computing importances greatly limits the impact of importance sampling. We propose a robust, approximate importance sampling procedure (RAIS) for stochastic gradient de- scent. By approximating the ideal sampling distribution using robust optimization, RAIS provides much of the benefit of exact importance sampling with drastically reduced overhead. Empirically, we find RAIS-SGD and standard SGD follow similar learning curves, but RAIS moves faster through these paths, achieving speed-ups of at least 20% and sometimes much more.

Collaborative Learning for Deep Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #59

Guocong Song · Wei Chai

We introduce collaborative learning in which multiple classifier heads of the same network are simultaneously trained on the same training data to improve generalization and robustness to label noise with no extra inference cost. It acquires the strengths from auxiliary training, multi-task learning and knowledge distillation. There are two important mechanisms involved in collaborative learning. First, the consensus of multiple views from different classifier heads on the same example provides supplementary information as well as regularization to each classifier, thereby improving generalization. Second, intermediate-level representation (ILR) sharing with backpropagation rescaling aggregates the gradient flows from all heads, which not only reduces training computational complexity, but also facilitates supervision to the shared layers. The empirical results on CIFAR and ImageNet datasets demonstrate that deep neural networks learned as a group in a collaborative way significantly reduce the generalization error and increase the robustness to label noise.

A Linear Speedup Analysis of Distributed Deep Learning with Sparse and Quantized Communication

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #60

Peng Jiang · Gagan Agrawal

The large communication overhead has imposed a bottleneck on the performance of distributed Stochastic Gradient Descent (SGD) for training deep neural networks. Previous works have demonstrated the potential of using gradient sparsification and quantization to reduce the communication cost. However, there is still a lack of understanding about how sparse and quantized communication affects the convergence rate of the training algorithm. In this paper, we study the convergence rate of distributed SGD for non-convex optimization with two communication reducing strategies: sparse parameter averaging and gradient quantization. We show that $O(1/\sqrt{MK})$ convergence rate can be achieved if the sparsification and quantization hyperparameters are configured properly. We also propose a strategy called periodic quantized averaging (PQASGD) that further reduces the communication cost while preserving the $O(1/\sqrt{MK})$ convergence rate. Our evaluation validates our theoretical results and shows that our PQASGD can converge as fast as full-communication SGD with only $3\%-5\%$ communication data size.

Bayesian Distributed Stochastic Gradient Descent

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #61

Michael Teng · Frank Wood

We introduce Bayesian distributed stochastic gradient descent (BDSGD), a high-throughput algorithm for training deep neural networks on parallel clusters. This algorithm uses amortized inference in a deep generative model to perform joint posterior predictive inference of mini-batch gradient computation times in a compute cluster specific manner. Specifically, our algorithm mitigates the straggler effect in synchronous, gradient-based optimization by choosing an optimal cutoff beyond which mini-batch gradient messages from slow workers are ignored. In our experiments, we show that eagerly discarding the mini-batch gradient computations of stragglers not only increases throughput but actually increases the overall rate of convergence as a function of wall-clock time by virtue of eliminating idleness. The principal novel contribution and finding of this work goes beyond this by demonstrating that using the predicted run-times from a generative model of cluster worker performance improves substantially over the static-cutoff prior art, leading to reduced deep neural net training times on large computer clusters.

Regularizing by the Variance of the Activations' Sample-Variances

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #62

Etai Littwin · Lior Wolf

Normalization techniques play an important role in supporting efficient and often more effective training of deep neural networks. While conventional methods explicitly normalize the activations, we suggest to add a loss term instead. This new loss term encourages the variance of the activations to be stable and not vary from one random mini-batch to the next. As we prove, this encourages the activations to be distributed around a few distinct modes. We also show that if the inputs are from a mixture of two Gaussians, the new loss would either join the two together, or separate between them optimally in the LDA sense, depending on the prior probabilities. Finally, we are able to link the new regularization term to the batchnorm method, which provides it with a regularization perspective. Our experiments demonstrate an improvement in accuracy over the batchnorm technique for both CNNs and fully connected networks.

BML: A High-performance, Low-cost Gradient

Synchronization Algorithm for DML Training

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #63

Songtao Wang · Dan Li · Yang Cheng · Jinkun Geng · Yanshu Wang · Shuai Wang · Shu-Tao Xia · Jianping Wu

In distributed machine learning (DML), the network performance between machines significantly impacts the speed of iterative training. In this paper we propose BML, a new gradient synchronization algorithm with higher network performance and lower network cost than the current practice. BML runs on BCube network, instead of using the traditional Fat-Tree topology. BML algorithm is designed in such a way that, compared to the parameter server (PS) algorithm on a Fat-Tree network connecting the same number of server machines, BML achieves theoretically 1/k of the gradient synchronization time, with k/5 of switches (the typical number of k is 2~4). Experiments of LeNet-5 and VGG-19 benchmarks on a testbed with 9 dual-GPU servers show that, BML reduces the job completion time of DML training by up to 56.4%.

L4: Practical loss-based stepsize adaptation for deep learning

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #64

Michal Rolinek · Georg Martius

We propose a stepsize adaptation scheme for stochastic gradient descent. It operates directly with the loss function and rescales the gradient in order to make fixed predicted progress on the loss. We demonstrate its capabilities by conclusively improving the performance of Adam and Momentum optimizers. The enhanced optimizers with default hyperparameters consistently outperform their constant stepsize counterparts, even the best ones, without a measurable increase in computational cost. The performance is validated on multiple architectures including dense nets, CNNs, ResNets, and the recurrent Differential Neural Computer on classical datasets MNIST, fashion MNIST, CIFAR10 and others.

Synaptic Strength For Convolutional Neural Network

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #65

CHEN LIN · Zhao Zhong · Wu Wei · Junjie Yan

Convolutional Neural Networks(CNNs) are both computation and memory inten-sive which hindered their deployment in mobile devices. Inspired by the relevantconcept in neural science literature, we propose Synaptic Pruning: a data-drivenmethod to prune connections between input and output feature maps with a newlyproposed class of parameters called Synaptic Strength. Synaptic Strength is de-signed to capture the importance of a connection based on the amount of informa-tion it transports. Experiment results show the effectiveness of our approach. OnCIFAR-10, we prune connections for various CNN models with up to96%, whichresults in significant size reduction and computation saving. Further evaluation onImageNet demonstrates that synaptic pruning is able to discover efficient modelswhich is competitive to state-of-the-art compact CNNs such as MobileNet-V2andNasNet-Mobile. Our contribution is summarized as following: (1) We introduceSynaptic Strength, a new class of parameters for CNNs to indicate the importanceof each connections. (2) Our approach can prune various CNNs with high com-pression without compromising accuracy. (3) Further investigation shows, theproposed Synaptic Strength is a better indicator for kernel pruning compared withthe previous approach in both empirical result and theoretical analysis.

ChannelNets: Compact and Efficient Convolutional Neural Networks via Channel-Wise Convolutions

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #66

Hongyang Gao · Zhengyang Wang · Shuiwang Ji

Convolutional neural networks (CNNs) have shown great capability of solving various artificial intelligence tasks. However, the increasing model size has raised challenges in employing them in resource-limited applications. In this work, we propose to compress deep models by using channel-wise convolutions, which replace dense connections among feature maps with sparse ones in CNNs. Based on this novel operation, we build light-weight CNNs known as ChannelNets. ChannelNets use three instances of channel-wise convolutions; namely group channel-wise convolutions, depth-wise separable channel-wise convolutions, and the convolutional classification layer. Compared to prior CNNs designed for mobile devices, ChannelNets achieve a significant reduction in terms of the number of parameters and computational cost without loss in accuracy. Notably, our work represents the first attempt to compress the fully-connected classification layer, which usually accounts for about 25% of total parameters in compact CNNs. Experimental results on the ImageNet dataset demonstrate that ChannelNets achieve consistently better performance compared to prior methods.

Frequency-Domain Dynamic Pruning for Convolutional Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #67

Zhenhua Liu · Jizheng Xu · Xiulian Peng · Ruiqin Xiong

Deep convolutional neural networks have demonstrated their powerfulness in a variety of applications. However, the storage and computational requirements have largely restricted their further extensions on mobile devices. Recently, pruning of unimportant parameters has been used for both network compression and acceleration. Considering that there are spatial redundancy within most filters in a CNN, we propose a frequency-domain dynamic pruning scheme to exploit the spatial correlations. The frequency-domain coefficients are pruned dynamically in each iteration and different frequency bands are pruned discriminatively, given their different importance on accuracy. Experimental results demonstrate that the proposed scheme can outperform previous spatial-domain counterparts by a large margin. Specifically, it can achieve a compression ratio of 8.4x and a theoretical inference speed-up of 9.2x for ResNet-110, while the accuracy is even better than the reference model on CIFAR-110.

TETRIS: TilE-matching the TRemendous Irregular Sparsity

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #68

Yu Ji · Ling Liang · Lei Deng · Youyang Zhang · Youhui Zhang · Yuan Xie

Compressing neural networks by pruning weights with small magnitudes can significantly reduce the computation and storage cost. Although pruning makes the model smaller, it is difficult to get practical speedup in modern computing platforms such as CPU and GPU due to the irregularity. Structural pruning has attract a lot of research interest to make sparsity hardware-friendly. Increasing the sparsity granularity can lead to better hardware utilization, but it will compromise the sparsity for maintaining accuracy. In this work, we propose a novel method, TETRIS, to achieve both better hardware utilization and higher sparsity. Just like a tile-matching game, we cluster the irregularly distributed weights with small value into structured groups by reordering the input/output dimension and structurally prune them. Results show that it can achieve comparable sparsity with the irregular element-wise pruning and demonstrate negligible accuracy loss. The experiments also shows ideal speedup, which is proportional to the sparsity, on GPU platforms. Our proposed method provides a new solution toward algorithm and architecture co-optimization for accuracy-efficiency trade-off.

Heterogeneous Bitwidth Binarization in Convolutional Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #69

Joshua Fromm · Shwetak Patel · Matthai Philipose

Recent work has shown that fast, compact low-bitwidth neural networks can be surprisingly accurate. These networks use homogeneous binarization: all parameters in each layer or (more commonly) the whole model have the same low bitwidth (e.g., 2 bits). However, modern hardware allows efficient designs where each arithmetic instruction can have a custom bitwidth, motivating heterogeneous binarization, where every parameter in the network may have a different bitwidth. In this paper, we show that it is feasible and useful to select bitwidths at the parameter granularity during training. For instance a heterogeneously quantized version of modern networks such as AlexNet and MobileNet, with the right mix of 1-, 2- and 3-bit parameters that average to just 1.4 bits can equal the accuracy of homogeneous 2-bit versions of these networks. Further, we provide analyses to show that the heterogeneously binarized systems yield FPGA- and ASIC-based implementations that are correspondingly more efficient in both circuit area and energy efficiency than their homogeneous counterparts.

HitNet: Hybrid Ternary Recurrent Neural Network

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #70

Peiqi Wang · Xinfeng Xie · Lei Deng · Guoqi Li · Dongsheng Wang · Yuan Xie

Quantization is a promising technique to reduce the model size, memory footprint, and massive computation operations of recurrent neural networks (RNNs) for embedded devices with limited resources. Although extreme low-bit quantization has achieved impressive success on convolutional neural networks, it still suffers from huge accuracy degradation on RNNs with the same low-bit precision. In this paper, we first investigate the accuracy degradation on RNN models under different quantization schemes, and the distribution of tensor values in the full precision model. Our observation reveals that due to the difference between the distributions of weights and activations, different quantization methods are suitable for different parts of models. Based on our observation, we propose HitNet, a hybrid ternary recurrent neural network, which bridges the accuracy gap between the full precision model and the quantized model. In HitNet, we develop a hybrid quantization method to quantize weights and activations. Moreover, we introduce a sloping factor motivated by prior work on Boltzmann machine to activation functions, further closing the accuracy gap between the full precision model and the quantized model. Overall, our HitNet can quantize RNN models into ternary values, {-1, 0, 1}, outperforming the state-of-the-art quantization methods on RNN models significantly. We test it on typical RNN models, such as Long-Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), on which the results outperform previous work significantly. For example, we improve the perplexity per word (PPW) of a ternary LSTM on Penn Tree Bank (PTB) corpus from 126 (the state-of-the-art result to the best of our knowledge) to 110.3 with a full precision model in 97.2, and a ternary GRU from 142 to 113.5 with a full precision model in 102.7.

A General Method for Amortizing Variational Filtering

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #71

Joseph Marino · Milan Cvitkovic · Yisong Yue

We introduce the variational filtering EM algorithm, a simple, general-purpose method for performing variational inference in dynamical latent variable models using information from only past and present variables, i.e. filtering. The algorithm is derived from the variational objective in the filtering setting and consists of an optimization procedure at each time step. By performing each inference optimization procedure with an iterative amortized inference model, we obtain a computationally efficient implementation of the algorithm, which we call amortized variational filtering. We present experiments demonstrating that this general-purpose method improves inference performance across several recent deep dynamical latent variable models.

Multiple Instance Learning for Efficient Sequential Data Classification on Resource-constrained Devices

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #72

Don Dennis · Chirag Pabbaraju · Harsha Vardhan Simhadri · Prateek Jain

We study the problem of fast and efficient classification of sequential data (such as time-series) on tiny devices, which is critical for various IoT related applications like audio keyword detection or gesture detection. Such tasks are cast as a standard classification task by sliding windows over the data stream to construct data points. Deploying such classification modules on tiny devices is challenging as predictions over sliding windows of data need to be invoked continuously at a high frequency. Each such predictor instance in itself is expensive as it evaluates large models over long windows of data. In this paper, we address this challenge by exploiting the following two observations about classification tasks arising in typical IoT related applications: (a) the "signature" of a particular class (e.g. an audio keyword) typically occupies a small fraction of the overall data, and (b) class signatures tend to be discernible early on in the data. We propose a method, EMI-RNN, that exploits these observations by using a multiple instance learning formulation along with an early prediction technique to learn a model that achieves better accuracy compared to baseline models, while simultaneously reducing computation by a large fraction. For instance, on a gesture detection benchmark [25], EMI-RNN improves standard LSTM model’s accuracy by up to 1% while requiring 72x less computation. This enables us to deploy such models for continuous real-time prediction on a small device such as Raspberry Pi0 and Arduino variants, a task that the baseline LSTM could not achieve. Finally, we also provide an analysis of our multiple instance learning algorithm in a simple setting and show that the proposed algorithm converges to the global optima at a linear rate, one of the first such result in this domain. The code for EMI-RNN is available at: https://github.com/Microsoft/EdgeML/tree/master/tf/examples/EMI-RNN

Navigating with Graph Representations for Fast and Scalable Decoding of Neural Language Models

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #73

Minjia Zhang · Wenhan Wang · Xiaodong Liu · Jianfeng Gao · Yuxiong He

Neural language models (NLMs) have recently gained a renewed interest by achieving state-of-the-art performance across many natural language processing (NLP) tasks. However, NLMs are very computationally demanding largely due to the computational cost of the decoding process, which consists of a softmax layer over a large vocabulary.We observe that in the decoding of many NLP tasks, only the probabilities of the top-K hypotheses need to be calculated preciously and K is often much smaller than the vocabulary size. This paper proposes a novel softmax layer approximation algorithm, called Fast Graph Decoder (FGD), which quickly identifies, for a given context, a set of K words that are most likely to occur according to a NLM. We demonstrate that FGD reduces the decoding time by an order of magnitude while attaining close to the full softmax baseline accuracy on neural machine translation and language modeling tasks. We also prove the theoretical guarantee on the softmax approximation quality.

Representer Point Selection for Explaining Deep Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #74

Chih-Kuan Yeh · Joon Kim · Ian En-Hsu Yen · Pradeep Ravikumar

We propose to explain the predictions of a deep neural network, by pointing to the set of what we call representer points in the training set, for a given test point prediction. Specifically, we show that we can decompose the pre-activation prediction of a neural network into a linear combination of activations of training points, with the weights corresponding to what we call representer values, which thus capture the importance of that training point on the learned parameters of the network. But it provides a deeper understanding of the network than simply training point influence: with positive representer values corresponding to excitatory training points, and negative values corresponding to inhibitory points, which as we show provides considerably more insight. Our method is also much more scalable, allowing for real-time feedback in a manner not feasible with influence functions.

Interpreting Neural Network Judgments via Minimal, Stable, and Symbolic Corrections

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #75

Xin Zhang · Armando Solar-Lezama · Rishabh Singh

We present a new algorithm to generate minimal, stable, and symbolic corrections to an input that will cause a neural network with ReLU activations to change its output. We argue that such a correction is a useful way to provide feedback to a user when the network's output is different from a desired output. Our algorithm generates such a correction by solving a series of linear constraint satisfaction problems. The technique is evaluated on three neural network models: one predicting whether an applicant will pay a mortgage, one predicting whether a first-order theorem can be proved efficiently by a solver using certain heuristics, and the final one judging whether a drawing is an accurate rendition of a canonical drawing of a cat.

DropMax: Adaptive Variational Softmax

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #76

Hae Beom Lee · Juho Lee · Saehoon Kim · Eunho Yang · Sung Ju Hwang

We propose DropMax, a stochastic version of softmax classifier which at each iteration drops non-target classes according to dropout probabilities adaptively decided for each instance. Specifically, we overlay binary masking variables over class output probabilities, which are input-adaptively learned via variational inference. This stochastic regularization has an effect of building an ensemble classifier out of exponentially many classifiers with different decision boundaries. Moreover, the learning of dropout rates for non-target classes on each instance allows the classifier to focus more on classification against the most confusing classes. We validate our model on multiple public datasets for classification, on which it obtains significantly improved accuracy over the regular softmax classifier and other baselines. Further analysis of the learned dropout probabilities shows that our model indeed selects confusing classes more often when it performs classification.

Out-of-Distribution Detection using Multiple Semantic Label Representations

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #77

Gabi Shalev · Yossi Adi · Joseph Keshet

Deep Neural Networks are powerful models that attained remarkable results on a variety of tasks.

These models are shown to be extremely efficient when training and test data are drawn from the same distribution. However, it is not clear how a network will act when it is fed with an out-of-distribution example. In this work, we consider the problem of out-of-distribution detection in neural networks. We propose to use multiple semantic dense representations instead of sparse representation as the target label. Specifically, we propose to use several word representations obtained from different corpora or architectures as target labels. We evaluated the proposed model on computer vision, and speech commands detection tasks and compared it to previous methods. Results suggest that our method compares favorably with previous work. Besides, we present the efficiency of our approach for detecting wrongly classified and adversarial examples.

End-to-end Symmetry Preserving Inter-atomic Potential Energy Model for Finite and Extended Systems

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #78

Linfeng Zhang · Jiequn Han · Han Wang · Wissam Saidi · Roberto Car · Weinan E

Machine learning models are changing the paradigm of molecular modeling, which is a fundamental tool for material science, chemistry, and computational biology. Of particular interest is the inter-atomic potential energy surface (PES). Here we develop Deep Potential - Smooth Edition (DeepPot-SE), an end-to-end machine learning-based PES model, which is able to efficiently represent the PES for a wide variety of systems with the accuracy of ab initio quantum mechanics models. By construction, DeepPot-SE is extensive and continuously differentiable, scales linearly with system size, and preserves all the natural symmetries of the system. Further, we show that DeepPot-SE describes finite and extended systems including organic molecules, metals, semiconductors, and insulators with high fidelity.

Deep Predictive Coding Network with Local Recurrent Processing for Object Recognition

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #79

Kuan Han · Haiguang Wen · Yizhen Zhang · Di Fu · Eugenio Culurciello · Zhongming Liu

Inspired by "predictive coding" - a theory in neuroscience, we develop a bi-directional and dynamic neural network with local recurrent processing, namely predictive coding network (PCN). Unlike feedforward-only convolutional neural networks, PCN includes both feedback connections, which carry top-down predictions, and feedforward connections, which carry bottom-up errors of prediction. Feedback and feedforward connections enable adjacent layers to interact locally and recurrently to refine representations towards minimization of layer-wise prediction errors. When unfolded over time, the recurrent processing gives rise to an increasingly deeper hierarchy of nonlinear transformation, allowing a shallow network to dynamically extend itself into an arbitrarily deep network. We train and test PCN for image classification with SVHN, CIFAR and ImageNet datasets. Despite notably fewer layers and parameters, PCN achieves competitive performance compared to classical and state-of-the-art models. Further analysis shows that the internal representations in PCN converge over time and yield increasingly better accuracy in object recognition. Errors of top-down prediction also reveal visual saliency or bottom-up attention.

SLAYER: Spike Layer Error Reassignment in Time

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #80

Sumit Bam Shrestha · Garrick Orchard

Configuring deep Spiking Neural Networks (SNNs) is an exciting research avenue for low power spike event based computation. However, the spike generation function is non-differentiable and therefore not directly compatible with the standard error backpropagation algorithm. In this paper, we introduce a new general backpropagation mechanism for learning synaptic weights and axonal delays which overcomes the problem of non-differentiability of the spike function and uses a temporal credit assignment policy for backpropagating error to preceding layers. We describe and release a GPU accelerated software implementation of our method which allows training both fully connected and convolutional neural network (CNN) architectures. Using our software, we compare our method against existing SNN based learning approaches and standard ANN to SNN conversion techniques and show that our method achieves state of the art performance for an SNN on the MNIST, NMNIST, DVS Gesture, and TIDIGITS datasets.

DeepPINK: reproducible feature selection in deep neural networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #81

Yang Lu · Yingying Fan · Jinchi Lv · William Stafford Noble

Deep learning has become increasingly popular in both supervised and unsupervised machine learning thanks to its outstanding empirical performance. However, because of their intrinsic complexity, most deep learning methods are largely treated as black box tools with little interpretability. Even though recent attempts have been made to facilitate the interpretability of deep neural networks (DNNs), existing methods are susceptible to noise and lack of robustness. Therefore, scientists are justifiably cautious about the reproducibility of the discoveries, which is often related to the interpretability of the underlying statistical models. In this paper, we describe a method to increase the interpretability and reproducibility of DNNs by incorporating the idea of feature selection with controlled error rate. By designing a new DNN architecture and integrating it with the recently proposed knockoffs framework, we perform feature selection with a controlled error rate, while maintaining high power. This new method, DeepPINK (Deep feature selection using Paired-Input Nonlinear Knockoffs), is applied to both simulated and real data sets to demonstrate its empirical utility.

Learning long-range spatial dependencies with horizontal gated recurrent units

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #82

Drew Linsley · Junkyung Kim · Vijay Veerabadran · Charles Windolf · Thomas Serre

Progress in deep learning has spawned great successes in many engineering applications. As a prime example, convolutional neural networks, a type of feedforward neural networks, are now approaching -- and sometimes even surpassing -- human accuracy on a variety of visual recognition tasks. Here, however, we show that these neural networks and their recent extensions struggle in recognition tasks where co-dependent visual features must be detected over long spatial ranges. We introduce a visual challenge, Pathfinder, and describe a novel recurrent neural network architecture called the horizontal gated recurrent unit (hGRU) to learn intrinsic horizontal connections -- both within and across feature columns. We demonstrate that a single hGRU layer matches or outperforms all tested feedforward hierarchical baselines including state-of-the-art architectures with orders of magnitude more parameters.

Neural Interaction Transparency (NIT): Disentangling Learned Interactions for Improved Interpretability

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #83

Michael Tsang · Hanpeng Liu · Sanjay Purushotham · Pavankumar Murali · Yan Liu

Neural networks are known to model statistical interactions, but they entangle the interactions at intermediate hidden layers for shared representation learning. We propose a framework, Neural Interaction Transparency (NIT), that disentangles the shared learning across different interactions to obtain their intrinsic lower-order and interpretable structure. This is done through a novel regularizer that directly penalizes interaction order. We show that disentangling interactions reduces a feedforward neural network to a generalized additive model with interactions, which can lead to transparent models that perform comparably to the state-of-the-art models. NIT is also flexible and efficient; it can learn generalized additive models with maximum K-order interactions by training only $O(1)$ models.

A Bridging Framework for Model Optimization and Deep Propagation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #84

Risheng Liu · Shichao Cheng · xiaokun liu · Long Ma · Xin Fan · Zhongxuan Luo

Optimizing task-related mathematical model is one of the most fundamental methodologies in statistic and learning areas. However, generally designed schematic iterations may hard to investigate complex data distributions in real-world applications. Recently, training deep propagations (i.e., networks) has gained promising performance in some particular tasks. Unfortunately, existing networks are often built in heuristic manners, thus lack of principled interpretations and solid theoretical supports. In this work, we provide a new paradigm, named Propagation and Optimization based Deep Model (PODM), to bridge the gaps between these different mechanisms (i.e., model optimization and deep propagation). On the one hand, we utilize PODM as a deeply trained solver for model optimization. Different from these existing network based iterations, which often lack theoretical investigations, we provide strict convergence analysis for PODM in the challenging nonconvex and nonsmooth scenarios. On the other hand, by relaxing the model constraints and performing end-to-end training, we also develop a PODM based strategy to integrate domain knowledge (formulated as models) and real data distributions (learned by networks), resulting in a generic ensemble framework for challenging real-world applications. Extensive experiments verify our theoretical results and demonstrate the superiority of PODM against these state-of-the-art approaches.

The Importance of Sampling inMeta-Reinforcement Learning

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #85

Bradly Stadie · Ge Yang · Rein Houthooft · Peter Chen · Yan Duan · Yuhuai Wu · Pieter Abbeel · Ilya Sutskever

We interpret meta-reinforcement learning as the problem of learning how to quickly find a good sampling distribution in a new environment. This interpretation leads to the development of two new meta-reinforcement learning algorithms: E-MAML and E-RL^2. Results are presented on a new environment we call `Krazy World': a difficult high-dimensional gridworld which is designed to highlight the importance of correctly differentiating through sampling distributions in meta-reinforcement learning. Further results are presented on a set of maze environments. We show E-MAML and E-RL^2 deliver better performance than baseline algorithms on both tasks.

Latent Alignment and Variational Attention

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #86

Yuntian Deng · Yoon Kim · Justin Chiu · Demi Guo · Alexander Rush

Neural attention has become central to many state-of-the-art models in natural language processing and related domains. Attention networks are an easy-to-train and effective method for softly simulating alignment; however, the approach does not marginalize over latent alignments in a probabilistic sense. This property makes it difficult to compare attention to other alignment approaches, to compose it with probabilistic models, and to perform posterior inference conditioned on observed data. A related latent approach, hard attention, fixes these issues, but is generally harder to train and less accurate. This work considers variational attention networks, alternatives to soft and hard attention for learning latent variable alignment models, with tighter approximation bounds based on amortized variational inference. We further propose methods for reducing the variance of gradients to make these approaches computationally feasible. Experiments show that for machine translation and visual question answering, inefficient exact latent variable models outperform standard neural attention, but these gains go away when using hard attention based training. On the other hand, variational attention retains most of the performance gain but with training speed comparable to neural attention.

Variational Memory Encoder-Decoder

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #87

Hung Le · Truyen Tran · Thin Nguyen · Svetha Venkatesh

Introducing variability while maintaining coherence is a core task in learning to generate utterances in conversation. Standard neural encoder-decoder models and their extensions using conditional variational autoencoder often result in either trivial or digressive responses. To overcome this, we explore a novel approach that injects variability into neural encoder-decoder via the use of external memory as a mixture model, namely Variational Memory Encoder-Decoder (VMED). By associating each memory read with a mode in the latent mixture distribution at each timestep, our model can capture the variability observed in sequential data such as natural conversations. We empirically compare the proposed model against other recent approaches on various conversational datasets. The results show that VMED consistently achieves significant improvement over others in both metric-based and qualitative evaluations.

Relational recurrent neural networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #88

Adam Santoro · Ryan Faulkner · David Raposo · Jack Rae · Mike Chrzanowski · Theophane Weber · Daan Wierstra · Oriol Vinyals · Razvan Pascanu · Timothy Lillicrap

Memory-based neural networks model temporal data by leveraging an ability to remember information for long periods. It is unclear, however, whether they also have an ability to perform complex relational reasoning with the information they remember. Here, we first confirm our intuitions that standard memory architectures may struggle at tasks that heavily involve an understanding of the ways in which entities are connected -- i.e., tasks involving relational reasoning. We then improve upon these deficits by using a new memory module -- a Relational Memory Core (RMC) -- which employs multi-head dot product attention to allow memories to interact. Finally, we test the RMC on a suite of tasks that may profit from more capable relational reasoning across sequential information, and show large gains in RL domains (BoxWorld & Mini PacMan), program evaluation, and language modeling, achieving state-of-the-art results on the WikiText-103, Project Gutenberg, and GigaWord datasets.

Learning to Reason with Third Order Tensor Products

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #89

Imanol Schlag · Jürgen Schmidhuber

We combine Recurrent Neural Networks with Tensor Product Representations to learn combinatorial representations of sequential data. This improves symbolic interpretation and systematic generalisation. Our architecture is trained end-to-end through gradient descent on a variety of simple natural language reasoning tasks, significantly outperforming the latest state-of-the-art models in single-task and all-tasks settings. We also augment a subset of the data such that training and test data exhibit large systematic differences and show that our approach generalises better than the previous state-of-the-art.

Reversible Recurrent Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #90

Matthew MacKay · Paul Vicol · Jimmy Ba · Roger Grosse

Recurrent neural networks (RNNs) provide state-of-the-art performance in processing sequential data but are memory intensive to train, limiting the flexibility of RNN models which can be trained. Reversible RNNs---RNNs for which the hidden-to-hidden transition can be reversed---offer a path to reduce the memory requirements of training, as hidden states need not be stored and instead can be recomputed during backpropagation. We first show that perfectly reversible RNNs, which require no storage of the hidden activations, are fundamentally limited because they cannot forget information from their hidden state. We then provide a scheme for storing a small number of bits in order to allow perfect reversal with forgetting. Our method achieves comparable performance to traditional models while reducing the activation memory cost by a factor of 10--15. We extend our technique to attention-based sequence-to-sequence models, where it maintains performance while reducing activation memory cost by a factor of 5--10 in the encoder, and a factor of 10--15 in the decoder.

Breaking the Activation Function Bottleneck through Adaptive Parameterization

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #91

Sebastian Flennerhag · Hujun Yin · John Keane · Mark Elliot

Standard neural network architectures are non-linear only by virtue of a simple element-wise activation function, making them both brittle and excessively large. In this paper, we consider methods for making the feed-forward layer more flexible while preserving its basic structure. We develop simple drop-in replacements that learn to adapt their parameterization conditional on the input, thereby increasing statistical efficiency significantly. We present an adaptive LSTM that advances the state of the art for the Penn Treebank and Wikitext-2 word-modeling tasks while using fewer parameters and converging in half as many iterations.

BRITS: Bidirectional Recurrent Imputation for Time Series

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #92

Wei Cao · Dong Wang · Jian Li · Hao Zhou · Lei Li · Yitan Li

Time series are widely used as signals in many classification/regression tasks. It is ubiquitous that time series contains many missing values. Given multiple correlated time series data, how to fill in missing values and to predict their class labels? Existing imputation methods often impose strong assumptions of the underlying data generating process, such as linear dynamics in the state space. In this paper, we propose BRITS, a novel method based on recurrent neural networks for missing value imputation in time series data. Our proposed method directly learns the missing values in a bidirectional recurrent dynamical system, without any specific assumption. The imputed values are treated as variables of RNN graph and can be effectively updated during the backpropagation. BRITS has three advantages: (a) it can handle multiple correlated missing values in time series; (b) it generalizes to time series with nonlinear dynamics underlying; (c) it provides a data-driven imputation procedure and applies to general settings with missing data. We evaluate our model on three real-world datasets, including an air quality dataset, a health-care data, and a localization data for human activity. Experiments show that our model outperforms the state-of-the-art methods in both imputation and classification/regression accuracies.

Complex Gated Recurrent Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #93

Moritz Wolter · Angela Yao

Complex numbers have long been favoured for digital signal processing, yet complex representations rarely appear in deep learning architectures. RNNs, widely used to process time series and sequence information, could greatly benefit from complex representations. We present a novel complex gated recurrent cell, which is a hybrid cell combining complex-valued and norm-preserving state transitions with a gating mechanism. The resulting RNN exhibits excellent stability and convergence properties and performs competitively on the synthetic memory and adding task, as well as on the real-world tasks of human motion prediction.

Middle-Out Decoding

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #94

Shikib Mehri · Leonid Sigal

Despite being virtually ubiquitous, sequence-to-sequence models are challenged by their lack of diversity and inability to be externally controlled. In this paper, we speculate that a fundamental shortcoming of sequence generation models is that the decoding is done strictly from left-to-right, meaning that outputs values generated earlier have a profound effect on those generated later. To address this issue, we propose a novel middle-out decoder architecture that begins from an initial middle-word and simultaneously expands the sequence in both directions. To facilitate information flow and maintain consistent decoding, we introduce a dual self-attention mechanism that allows us to model complex dependencies between the outputs. We illustrate the performance of our model on the task of video captioning, as well as a synthetic sequence de-noising task. Our middle-out decoder achieves significant improvements on de-noising and competitive performance in the task of video captioning, while quantifiably improving the caption diversity. Furthermore, we perform a qualitative analysis that demonstrates our ability to effectively control the generation process of our decoder.

Recurrently Controlled Recurrent Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #95

Yi Tay · Anh Tuan Luu · Siu Cheung Hui

Recurrent neural networks (RNNs) such as long short-term memory and gated recurrent units are pivotal building blocks across a broad spectrum of sequence modeling problems. This paper proposes a recurrently controlled recurrent network (RCRN) for expressive and powerful sequence encoding. More concretely, the key idea behind our approach is to learn the recurrent gating functions using recurrent networks. Our architecture is split into two components - a controller cell and a listener cell whereby the recurrent controller actively influences the compositionality of the listener cell. We conduct extensive experiments on a myriad of tasks in the NLP domain such as sentiment analysis (SST, IMDb, Amazon reviews, etc.), question classification (TREC), entailment classification (SNLI, SciTail), answer selection (WikiQA, TrecQA) and reading comprehension (NarrativeQA). Across all 26 datasets, our results demonstrate that RCRN not only consistently outperforms BiLSTMs but also stacked BiLSTMs, suggesting that our controller architecture might be a suitable replacement for the widely adopted stacked architecture. Additionally, RCRN achieves state-of-the-art results on several well-established datasets.

Tree-to-tree Neural Networks for Program Translation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #96

Xinyun Chen · Chang Liu · Dawn Song

Program translation is an important tool to migrate legacy code in one language into an ecosystem built in a different language. In this work, we are the first to employ deep neural networks toward tackling this problem. We observe that program translation is a modular procedure, in which a sub-tree of the source tree is translated into the corresponding target sub-tree at each step. To capture this intuition, we design a tree-to-tree neural network to translate a source tree into a target one. Meanwhile, we develop an attention mechanism for the tree-to-tree model, so that when the decoder expands one non-terminal in the target tree, the attention mechanism locates the corresponding sub-tree in the source tree to guide the expansion of the decoder. We evaluate the program translation capability of our tree-to-tree model against several state-of-the-art approaches. Compared against other neural translation models, we observe that our approach is consistently better than the baselines with a margin of up to 15 points. Further, our approach can improve the previous state-of-the-art program translation approaches by a margin of 20 points on the translation of real-world projects.

HOUDINI: Lifelong Learning as Program Synthesis

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #97

Lazar Valkov · Dipak Chaudhari · Akash Srivastava · Charles Sutton · Swarat Chaudhuri

We present a neurosymbolic framework for the lifelong learning of algorithmic tasks that mix perception and procedural reasoning. Reusing high-level concepts across domains and learning complex procedures are key challenges in lifelong learning. We show that a program synthesis approach that combines gradient descent with combinatorial search over programs can be a more effective response to these challenges than purely neural methods. Our framework, called HOUDINI, represents neural networks as strongly typed, differentiable functional programs that use symbolic higher-order combinators to compose a library of neural functions. Our learning algorithm consists of: (1) a symbolic program synthesizer that performs a type-directed search over parameterized programs, and decides on the library functions to reuse, and the architectures to combine them, while learning a sequence of tasks; and (2) a neural module that trains these programs using stochastic gradient descent. We evaluate HOUDINI on three benchmarks that combine perception with the algorithmic tasks of counting, summing, and shortest-path computation. Our experiments show that HOUDINI transfers high-level concepts more effectively than traditional transfer learning and progressive neural networks, and that the typed representation of networks significantly accelerates the search.

Neural Guided Constraint Logic Programming for Program Synthesis

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #98

Lisa Zhang · Gregory Rosenblatt · Ethan Fetaya · Renjie Liao · William Byrd · Matthew Might · Raquel Urtasun · Richard Zemel

Synthesizing programs using example input/outputs is a classic problem in artificial intelligence. We present a method for solving Programming By Example (PBE) problems by using a neural model to guide the search of a constraint logic programming system called miniKanren. Crucially, the neural model uses miniKanren's internal representation as input; miniKanren represents a PBE problem as recursive constraints imposed by the provided examples. We explore Recurrent Neural Network and Graph Neural Network models. We contribute a modified miniKanren, drivable by an external agent, available at https://github.com/xuexue/neuralkanren. We show that our neural-guided approach using constraints can synthesize programs faster in many cases, and importantly, can generalize to larger problems.

Embedding Logical Queries on Knowledge Graphs

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #99

Will Hamilton · Payal Bajaj · Marinka Zitnik · Dan Jurafsky · Jure Leskovec

Learning low-dimensional embeddings of knowledge graphs is a powerful approach used to predict unobserved or missing edges between entities. However, an open challenge in this area is developing techniques that can go beyond simple edge prediction and handle more complex logical queries, which might involve multiple unobserved edges, entities, and variables. For instance, given an incomplete biological knowledge graph, we might want to predict "em what drugs are likely to target proteins involved with both diseases X and Y?" -- a query that requires reasoning about all possible proteins that might interact with diseases X and Y. Here we introduce a framework to efficiently make predictions about conjunctive logical queries -- a flexible but tractable subset of first-order logic -- on incomplete knowledge graphs. In our approach, we embed graph nodes in a low-dimensional space and represent logical operators as learned geometric operations (e.g., translation, rotation) in this embedding space. By performing logical operations within a low-dimensional embedding space, our approach achieves a time complexity that is linear in the number of query variables, compared to the exponential complexity required by a naive enumeration-based approach. We demonstrate the utility of this framework in two application studies on real-world datasets with millions of relations: predicting logical relationships in a network of drug-gene-disease interactions and in a graph-based representation of social interactions derived from a popular web forum.

Expanding Holographic Embeddings for Knowledge Completion

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #100

Yexiang Xue · Yang Yuan · Zhitian Xu · Ashish Sabharwal

Neural models operating over structured spaces such as knowledge graphs require a continuous embedding of the discrete elements of this space (such as entities) as well as the relationships between them. Relational embeddings with high expressivity, however, have high model complexity, making them computationally difficult to train. We propose a new family of embeddings for knowledge graphs that interpolate between a method with high model complexity and one, namely Holographic embeddings (HolE), with low dimensionality and high training efficiency. This interpolation, termed HolEx, is achieved by concatenating several linearly perturbed copies of original HolE. We formally characterize the number of perturbed copies needed to provably recover the full entity-entity or entity-relation interaction matrix, leveraging ideas from Haar wavelets and compressed sensing. In practice, using just a handful of Haar-based or random perturbation vectors results in a much stronger knowledge completion system. On the Freebase FB15K dataset, HolEx outperforms originally reported HolE by 14.7\% on the HITS@10 metric, and the current path-based state-of-the-art method, PTransE, by 4\% (absolute).

On the Dimensionality of Word Embedding

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #101

Zi Yin · Yuanyuan Shen

In this paper, we provide a theoretical understanding of word embedding and its dimensionality. Motivated by the unitary-invariance of word embedding, we propose the Pairwise Inner Product (PIP) loss, a novel metric on the dissimilarity between word embeddings. Using techniques from matrix perturbation theory, we reveal a fundamental bias-variance trade-off in dimensionality selection for word embeddings. This bias-variance trade-off sheds light on many empirical observations which were previously unexplained, for example the existence of an optimal dimensionality. Moreover, new insights and discoveries, like when and how word embeddings are robust to over-fitting, are revealed. By optimizing over the bias-variance trade-off of the PIP loss, we can explicitly answer the open question of dimensionality selection for word embedding.

Flexible neural representation for physics prediction

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #102

Damian Mrowca · Chengxu Zhuang · Elias Wang · Nick Haber · Li Fei-Fei · Josh Tenenbaum · Daniel Yamins

Humans have a remarkable capacity to understand the physical dynamics of objects in their environment, flexibly capturing complex structures and interactions at multiple levels of detail. Inspired by this ability, we propose a hierarchical particle-based object representation that covers a wide variety of types of three-dimensional objects, including both arbitrary rigid geometrical shapes and deformable materials. We then describe the Hierarchical Relation Network (HRN), an end-to-end differentiable neural network based on hierarchical graph convolution, that learns to predict physical dynamics in this representation. Compared to other neural network baselines, the HRN accurately handles complex collisions and nonrigid deformations, generating plausible dynamics predictions at long time scales in novel settings, and scaling to large scene configurations. These results demonstrate an architecture with the potential to form the basis of next-generation physics predictors for use in computer vision, robotics, and quantitative cognitive science.

Content preserving text generation with attribute controls

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #103

Lajanugen Logeswaran · Honglak Lee · Samy Bengio

In this work, we address the problem of modifying textual attributes of sentences. Given an input sentence and a set of attribute labels, we attempt to generate sentences that are compatible with the conditioning information. To ensure that the model generates content compatible sentences, we introduce a reconstruction loss which interpolates between auto-encoding and back-translation loss components. We propose an adversarial loss to enforce generated samples to be attribute compatible and realistic. Through quantitative, qualitative and human evaluations we demonstrate that our model is capable of generating fluent sentences that better reflect the conditioning information compared to prior methods. We further demonstrate that the model is capable of simultaneously controlling multiple attributes.

Recurrent Relational Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #104

Rasmus Palm · Ulrich Paquet · Ole Winther

This paper is concerned with learning to solve tasks that require a chain of interde- pendent steps of relational inference, like answering complex questions about the relationships between objects, or solving puzzles where the smaller elements of a solution mutually constrain each other. We introduce the recurrent relational net- work, a general purpose module that operates on a graph representation of objects. As a generalization of Santoro et al. [2017]'s relational network, it can augment any neural network model with the capacity to do many-step relational reasoning. We achieve state of the art results on the bAbI textual question-answering dataset with the recurrent relational network, consistently solving 20/20 tasks. As bAbI is not particularly challenging from a relational reasoning point of view, we introduce Pretty-CLEVR, a new diagnostic dataset for relational reasoning. In the Pretty- CLEVR set-up, we can vary the question to control for the number of relational reasoning steps that are required to obtain the answer. Using Pretty-CLEVR, we probe the limitations of multi-layer perceptrons, relational and recurrent relational networks. Finally, we show how recurrent relational networks can learn to solve Sudoku puzzles from supervised training data, a challenging task requiring upwards of 64 steps of relational reasoning. We achieve state-of-the-art results amongst comparable methods by solving 96.6% of the hardest Sudoku puzzles.

GLoMo: Unsupervised Learning of Transferable Relational Graphs

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #105

Zhilin Yang · Jake Zhao · Bhuwan Dhingra · Kaiming He · William Cohen · Ruslan Salakhutdinov · Yann LeCun

Modern deep transfer learning approaches have mainly focused on learning generic feature vectors from one task that are transferable to other tasks, such as word embeddings in language and pretrained convolutional features in vision. However, these approaches usually transfer unary features and largely ignore more structured graphical representations. This work explores the possibility of learning generic latent relational graphs that capture dependencies between pairs of data units (e.g., words or pixels) from large-scale unlabeled data and transferring the graphs to downstream tasks. Our proposed transfer learning framework improves performance on various tasks including question answering, natural language inference, sentiment analysis, and image classification. We also show that the learned graphs are generic enough to be transferred to different embeddings on which the graphs have not been trained (including GloVe embeddings, ELMo embeddings, and task-specific RNN hidden units), or embedding-free units such as image pixels.

Predictive Uncertainty Estimation via Prior Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #106

Andrey Malinin · Mark Gales

Estimating how uncertain an AI system is in its predictions is important to improve the safety of such systems. Uncertainty in predictive can result from uncertainty in model parameters, irreducible \emph{data uncertainty} and uncertainty due to distributional mismatch between the test and training data distributions. Different actions might be taken depending on the source of the uncertainty so it is important to be able to distinguish between them. Recently, baseline tasks and metrics have been defined and several practical methods to estimate uncertainty developed. These methods, however, attempt to model uncertainty due to distributional mismatch either implicitly through \emph{model uncertainty} or as \emph{data uncertainty}. This work proposes a new framework for modeling predictive uncertainty called Prior Networks (PNs) which explicitly models \emph{distributional uncertainty}. PNs do this by parameterizing a prior distribution over predictive distributions. This work focuses on uncertainty for classification and evaluates PNs on the tasks of identifying out-of-distribution (OOD) samples and detecting misclassification on the MNIST and CIFAR-10 datasets, where they are found to outperform previous methods. Experiments on synthetic and MNIST and CIFAR-10 data show that unlike previous non-Bayesian methods PNs are able to distinguish between data and distributional uncertainty.

Adversarial Multiple Source Domain Adaptation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #107

Han Zhao · Shanghang Zhang · Guanhang Wu · José M. F. Moura · Joao P Costeira · Geoffrey Gordon

While domain adaptation has been actively researched, most algorithms focus on the single-source-single-target adaptation setting. In this paper we propose new generalization bounds and algorithms under both classification and regression settings for unsupervised multiple source domain adaptation. Our theoretical analysis naturally leads to an efficient learning strategy using adversarial neural networks: we show how to interpret it as learning feature representations that are invariant to the multiple domain shifts while still being discriminative for the learning task. To this end, we propose multisource domain adversarial networks (MDAN) that approach domain adaptation by optimizing task-adaptive generalization bounds. To demonstrate the effectiveness of MDAN, we conduct extensive experiments showing superior adaptation performance on both classification and regression problems: sentiment analysis, digit classification, and vehicle counting.

Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #108

Gamaleldin Elsayed · Shreya Shankar · Brian Cheung · Nicolas Papernot · Alexey Kurakin · Ian Goodfellow · Jascha Sohl-Dickstein

Machine learning models are vulnerable to adversarial examples: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

A Simple Cache Model for Image Recognition

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #109

Emin Orhan

Training large-scale image recognition models is computationally expensive. This raises the question of whether there might be simple ways to improve the test performance of an already trained model without having to re-train or fine-tune it with new data. Here, we show that, surprisingly, this is indeed possible. The key observation we make is that the layers of a deep network close to the output layer contain independent, easily extractable class-relevant information that is not contained in the output layer itself. We propose to extract this extra class-relevant information using a simple key-value cache memory to improve the classification performance of the model at test time. Our cache memory is directly inspired by a similar cache model previously proposed for language modeling (Grave et al., 2017). This cache component does not require any training or fine-tuning; it can be applied to any pre-trained model and, by properly setting only two hyper-parameters, leads to significant improvements in its classification performance. Improvements are observed across several architectures and datasets. In the cache component, using features extracted from layers close to the output (but not from the output layer itself) as keys leads to the largest improvements. Concatenating features from multiple layers to form keys can further improve performance over using single-layer features as keys. The cache component also has a regularizing effect, a simple consequence of which is that it substantially increases the robustness of models against adversarial attacks.

Co-teaching: Robust training of deep neural networks with extremely noisy labels

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #110

Bo Han · Quanming Yao · Xingrui Yu · Gang Niu · Miao Xu · Weihua Hu · Ivor Tsang · Masashi Sugiyama

Deep learning with noisy labels is practically challenging, as the capacity of deep models is so high that they can totally memorize these noisy labels sooner or later during training. Nonetheless, recent studies on the memorization effects of deep neural networks show that they would first memorize training data of clean labels and then those of noisy labels. Therefore in this paper, we propose a new deep learning paradigm called ''Co-teaching'' for combating with noisy labels. Namely, we train two deep neural networks simultaneously, and let them teach each other given every mini-batch: firstly, each network feeds forward all data and selects some data of possibly clean labels; secondly, two networks communicate with each other what data in this mini-batch should be used for training; finally, each network back propagates the data selected by its peer network and updates itself. Empirical results on noisy versions of MNIST, CIFAR-10 and CIFAR-100 demonstrate that Co-teaching is much superior to the state-of-the-art methods in the robustness of trained deep models.

Improved Network Robustness with Adversary Critic

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #111

Alexander Matyasko · Lap-Pui Chau

Ideally, what confuses neural network should be confusing to humans. However, recent experiments have shown that small, imperceptible perturbations can change the network prediction. To address this gap in perception, we propose a novel approach for learning robust classifier. Our main idea is: adversarial examples for the robust classifier should be indistinguishable from the regular data of the adversarial target. We formulate a problem of learning robust classifier in the framework of Generative Adversarial Networks (GAN), where the adversarial attack on classifier acts as a generator, and the critic network learns to distinguish between regular and adversarial images. The classifier cost is augmented with the objective that its adversarial examples should confuse the adversary critic. To improve the stability of the adversarial mapping, we introduce adversarial cycle-consistency constraint which ensures that the adversarial mapping of the adversarial examples is close to the original. In the experiments, we show the effectiveness of our defense. Our method surpasses in terms of robustness networks trained with adversarial training. Additionally, we verify in the experiments with human annotators on MTurk that adversarial examples are indeed visually confusing.

Unsupervised Learning of Object Landmarks through Conditional Image Generation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #112

Tomas Jakab · Ankush Gupta · Hakan Bilen · Andrea Vedaldi

We propose a method for learning landmark detectors for visual objects (such as the eyes and the nose in a face) without any manual supervision. We cast this as the problem of generating images that combine the appearance of the object as seen in a first example image with the geometry of the object as seen in a second example image, where the two examples differ by a viewpoint change and/or an object deformation. In order to factorize appearance and geometry, we introduce a tight bottleneck in the geometry-extraction process that selects and distils geometry-related features. Compared to standard image generation problems, which often use generative adversarial networks, our generation task is conditioned on both appearance and geometry and thus is significantly less ambiguous, to the point that adopting a simple perceptual loss formulation is sufficient. We demonstrate that our approach can learn object landmarks from synthetic image deformations or videos, all without manual supervision, while outperforming state-of-the-art unsupervised landmark detectors. We further show that our method is applicable to a large variety of datasets - faces, people, 3D objects, and digits - without any modifications.

Multi-Task Learning as Multi-Objective Optimization

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #113

Ozan Sener · Vladlen Koltun

In multi-task learning, multiple tasks are solved jointly, sharing inductive bias between them. Multi-task learning is inherently a multi-objective problem because different tasks may conflict, necessitating a trade-off. A common compromise is to optimize a proxy objective that minimizes a weighted linear combination of per-task losses. However, this workaround is only valid when the tasks do not compete, which is rarely the case. In this paper, we explicitly cast multi-task learning as multi-objective optimization, with the overall objective of finding a Pareto optimal solution. To this end, we use algorithms developed in the gradient-based multi-objective optimization literature. These algorithms are not directly applicable to large-scale learning problems since they scale poorly with the dimensionality of the gradients and the number of tasks. We therefore propose an upper bound for the multi-objective loss and show that it can be optimized efficiently. We further prove that optimizing this upper bound yields a Pareto optimal solution under realistic assumptions. We apply our method to a variety of multi-task deep learning problems including digit classification, scene understanding (joint semantic segmentation, instance segmentation, and depth estimation), and multi-label classification. Our method produces higher-performing models than recent multi-task learning formulations or per-task training.

Deep Anomaly Detection Using Geometric Transformations

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #114

Izhak Golan · Ran El-Yaniv

We consider the problem of anomaly detection in images, and present a new detection technique. Given a sample of images, all known to belong to a ``normal'' class (e.g., dogs), we show how to train a deep neural model that can detect out-of-distribution images (i.e., non-dog objects). The main idea behind our scheme is to train a multi-class model to discriminate between dozens of geometric transformations applied on all the given images. The auxiliary expertise learned by the model generates feature detectors that effectively identify, at test time, anomalous images based on the softmax activation statistics of the model when applied on transformed images. We present extensive experiments using the proposed detector, which indicate that our algorithm improves state-of-the-art methods by a wide margin.

Practical Deep Stereo (PDS): Toward applications-friendly deep stereo matching

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #115

Stepan Tulyakov · Anton Ivanov · François Fleuret

End-to-end deep-learning networks recently demonstrated extremely good performance for stereo matching. However, existing networks are difficult to use for practical applications since (1) they are memory-hungry and unable to process even modest-size images, (2) they have to be fully re-trained to handle a different disparity range. The Practical Deep Stereo (PDS) network that we propose addresses both issues: First, its architecture relies on novel bottleneck modules that drastically reduce the memory footprint in inference, and additional design choices allow to handle greater image size during training. This results in a model that leverages large image context to resolve matching ambiguities. Second, a novel sub-pixel cross-entropy loss combined with a MAP estimator make this network less sensitive to ambiguous matches, and applicable to any disparity range without re-training. We compare PDS to state-of-the-art methods published over the recent months, and demonstrate its superior performance on FlyingThings3D and KITTI sets.

VideoCapsuleNet: A Simplified Network for Action Detection

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #116

Kevin Duarte · Yogesh Rawat · Mubarak Shah

The recent advances in Deep Convolutional Neural Networks (DCNNs) have shown extremely good results for video human action classification, however, action detection is still a challenging problem. The current action detection approaches follow a complex pipeline which involves multiple tasks such as tube proposals, optical flow, and tube classification. In this work, we present a more elegant solution for action detection based on the recently developed capsule network. We propose a 3D capsule network for videos, called VideoCapsuleNet: a unified network for action detection which can jointly perform pixel-wise action segmentation along with action classification. The proposed network is a generalization of capsule network from 2D to 3D, which takes a sequence of video frames as input. The 3D generalization drastically increases the number of capsules in the network, making capsule routing computationally expensive. We introduce capsule-pooling in the convolutional capsule layer to address this issue and make the voting algorithm tractable. The routing-by-agreement in the network inherently models the action representations and various action characteristics are captured by the predicted capsules. This inspired us to utilize the capsules for action localization and the class-specific capsules predicted by the network are used to determine a pixel-wise localization of actions. The localization is further improved by parameterized skip connections with the convolutional capsule layers and the network is trained end-to-end with a classification as well as localization loss. The proposed network achieves state-of-the-art performance on multiple action detection datasets including UCF-Sports, J-HMDB, and UCF-101 (24 classes) with an impressive ~20% improvement on UCF-101 and ~15% improvement on J-HMDB in terms of v-mAP scores.

With Friends Like These, Who Needs Adversaries?

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #117

Saumya Jetley · Nicholas Lord · Philip Torr

The vulnerability of deep image classification networks to adversarial attack is now well known, but less well understood. Via a novel experimental analysis, we illustrate some facts about deep convolutional networks for image classification that shed new light on their behaviour and how it connects to the problem of adversaries. In short, the celebrated performance of these networks and their vulnerability to adversarial attack are simply two sides of the same coin: the input image-space directions along which the networks are most vulnerable to attack are the same directions which they use to achieve their classification performance in the first place. We develop this result in two main steps. The first uncovers the fact that classes tend to be associated with specific image-space directions. This is shown by an examination of the class-score outputs of nets as functions of 1D movements along these directions. This provides a novel perspective on the existence of universal adversarial perturbations. The second is a clear demonstration of the tight coupling between classification performance and vulnerability to adversarial attack within the spaces spanned by these directions. Thus, our analysis resolves the apparent contradiction between accuracy and vulnerability. It provides a new perspective on much of the prior art and reveals profound implications for efforts to construct neural nets that are both accurate and robust to adversarial attack.

Multi-Task Zipping via Layer-wise Neuron Sharing

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #118

Xiaoxi He · Zimu Zhou · Lothar Thiele

Future mobile devices are anticipated to perceive, understand and react to the world on their own by running multiple correlated deep neural networks on-device. Yet the complexity of these neural networks needs to be trimmed down both within-model and cross-model to fit in mobile storage and memory. Previous studies focus on squeezing the redundancy within a single neural network. In this work, we aim to reduce the redundancy across multiple models. We propose Multi-Task Zipping (MTZ), a framework to automatically merge correlated, pre-trained deep neural networks for cross-model compression. Central in MTZ is a layer-wise neuron sharing and incoming weight updating scheme that induces a minimal change in the error function. MTZ inherits information from each model and demands light retraining to re-boost the accuracy of individual tasks. Evaluations show that MTZ is able to fully merge the hidden layers of two VGG-16 networks with a 3.18% increase in the test error averaged on ImageNet and CelebA, or share 39.61% parameters between the two networks with <0.5% increase in the test errors for both tasks. The number of iterations to retrain the combined network is at least 17.8 times lower than that of training a single VGG-16 network. Moreover, experiments show that MTZ is also able to effectively merge multiple residual networks.

Learning Versatile Filters for Efficient Convolutional Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #119

Yunhe Wang · Chang Xu · Chunjing XU · Chao Xu · Dacheng Tao

This paper introduces versatile filters to construct efficient convolutional neural network. Considering the demands of efficient deep learning techniques running on cost-effective hardware, a number of methods have been developed to learn compact neural networks. Most of these works aim to slim down filters in different ways, e.g., investigating small, sparse or binarized filters. In contrast, we treat filters from an additive perspective. A series of secondary filters can be derived from a primary filter. These secondary filters all inherit in the primary filter without occupying more storage, but once been unfolded in computation they could significantly enhance the capability of the filter by integrating information extracted from different receptive fields. Besides spatial versatile filters, we additionally investigate versatile filters from the channel perspective. The new techniques are general to upgrade filters in existing CNNs. Experimental results on benchmark datasets and neural networks demonstrate that CNNs constructed with our versatile filters are able to achieve comparable accuracy as that of original filters, but require less memory and FLOPs.

Evolutionary Stochastic Gradient Descent for Optimization of Deep Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #120

Xiaodong Cui · Wei Zhang · Zoltán Tüske · Michael Picheny

We propose a population-based Evolutionary Stochastic Gradient Descent (ESGD) framework for optimizing deep neural networks. ESGD combines SGD and gradient-free evolutionary algorithms as complementary algorithms in one framework in which the optimization alternates between the SGD step and evolution step to improve the average fitness of the population. With a back-off strategy in the SGD step and an elitist strategy in the evolution step, it guarantees that the best fitness in the population will never degrade. In addition, individuals in the population optimized with various SGD-based optimizers using distinct hyper-parameters in the SGD step are considered as competing species in a coevolution setting such that the complementarity of the optimizers is also taken into account. The effectiveness of ESGD is demonstrated across multiple applications including speech recognition, image recognition and language modeling, using networks with a variety of deep architectures.

Structure-Aware Convolutional Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #121

Jianlong Chang · Jie Gu · Lingfeng Wang · GAOFENG MENG · SHIMING XIANG · Chunhong Pan

Convolutional neural networks (CNNs) are inherently subject to invariable filters that can only aggregate local inputs with the same topological structures. It causes that CNNs are allowed to manage data with Euclidean or grid-like structures (e.g., images), not ones with non-Euclidean or graph structures (e.g., traffic networks). To broaden the reach of CNNs, we develop structure-aware convolution to eliminate the invariance, yielding a unified mechanism of dealing with both Euclidean and non-Euclidean structured data. Technically, filters in the structure-aware convolution are generalized to univariate functions, which are capable of aggregating local inputs with diverse topological structures. Since infinite parameters are required to determine a univariate function, we parameterize these filters with numbered learnable parameters in the context of the function approximation theory. By replacing the classical convolution in CNNs with the structure-aware convolution, Structure-Aware Convolutional Neural Networks (SACNNs) are readily established. Extensive experiments on eleven datasets strongly evidence that SACNNs outperform current models on various machine learning tasks, including image classification and clustering, text categorization, skeleton-based action recognition, molecular activity detection, and taxi flow prediction.

Global Gated Mixture of Second-order Pooling for Improving Deep Convolutional Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #122

Qilong Wang · Zilin Gao · Jiangtao Xie · Wangmeng Zuo · Peihua Li

In most of existing deep convolutional neural networks (CNNs) for classification, global average (first-order) pooling (GAP) has become a standard module to summarize activations of the last convolution layer as final representation for prediction. Recent researches show integration of higher-order pooling (HOP) methods clearly improves performance of deep CNNs. However, both GAP and existing HOP methods assume unimodal distributions, which cannot fully capture statistics of convolutional activations, limiting representation ability of deep CNNs, especially for samples with complex contents. To overcome the above limitation, this paper proposes a global Gated

Mixture of Second-order Pooling (GM-SOP) method to further improve representation ability of deep CNNs. To this end, we introduce a sparsity-constrained gating mechanism and propose a novel parametric SOP as component of mixture model. Given a bank of SOP candidates, our method can adaptively choose Top-K (K > 1) candidates for each input sample through the sparsity-constrained gating module, and performs weighted sum of outputs of K selected candidates as representation of the sample. The proposed GM-SOP can flexibly accommodate a large number of personalized SOP candidates in an efficient way, leading to richer representations. The deep networks with our GM-SOP can be end-to-end trained, having potential to characterize complex, multi-modal distributions. The proposed method is evaluated on two large scale image benchmarks (i.e., downsampled ImageNet-1K and Places365), and experimental results show our GM-SOP is superior to its counterparts and achieves very competitive performance. The source code will be available at http://www.peihuali.org/GM-SOP.

Hybrid Macro/Micro Level Backpropagation for Training Deep Spiking Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #123

Yingyezhe Jin · Wenrui Zhang · Peng Li

Spiking neural networks (SNNs) are positioned to enable spatio-temporal information processing and ultra-low power event-driven neuromorphic hardware. However, SNNs are yet to reach the same performances of conventional deep artificial neural networks (ANNs), a long-standing challenge due to complex dynamics and non-differentiable spike events encountered in training. The existing SNN error backpropagation (BP) methods are limited in terms of scalability, lack of proper handling of spiking discontinuities, and/or mismatch between the rate-coded loss function and computed gradient. We present a hybrid macro/micro level backpropagation (HM2-BP) algorithm for training multi-layer SNNs. The temporal effects are precisely captured by the proposed spike-train level post-synaptic potential (S-PSP) at the microscopic level. The rate-coded errors are defined at the macroscopic level, computed and back-propagated across both macroscopic and microscopic levels. Different from existing BP methods, HM2-BP directly computes the gradient of the rate-coded loss function w.r.t tunable parameters. We evaluate the proposed HM2-BP algorithm by training deep fully connected and convolutional SNNs based on the static MNIST [14] and dynamic neuromorphic N-MNIST [26]. HM2-BP achieves an accuracy level of 99.49% and 98.88% for MNIST and N-MNIST, respectively, outperforming the best reported performances obtained from the existing SNN BP algorithms. Furthermore, the HM2-BP produces the highest accuracies based on SNNs for the EMNIST [3] dataset, and leads to high recognition accuracy for the 16-speaker spoken English letters of TI46 Corpus [16], a challenging patio-temporal speech recognition benchmark for which no prior success based on SNNs was reported. It also achieves competitive performances surpassing those of conventional deep learning models when dealing with asynchronous spiking streams.

Deep Neural Nets with Interpolating Function as Output Activation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #124

Bao Wang · Xiyang Luo · Zhen Li · Wei Zhu · Zuoqiang Shi · Stanley Osher

We replace the output layer of deep neural nets, typically the softmax function, by a novel interpolating function. And we propose end-to-end training and testing algorithms for this new architecture. Compared to classical neural nets with softmax function as output activation, the surrogate with interpolating function as output activation combines advantages of both deep and manifold learning. The new framework demonstrates the following major advantages: First, it is better applicable to the case with insufficient training data. Second, it significantly improves the generalization accuracy on a wide variety of networks. The algorithm is implemented in PyTorch, and the code is available at https://github.com/ BaoWangMath/DNN-DataDependentActivation.

Neural Edit Operations for Biological Sequences

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #125

Satoshi Koide · Keisuke Kawano · Takuro Kutsuna

The evolution of biological sequences, such as proteins or DNAs, is driven by the three basic edit operations: substitution, insertion, and deletion. Motivated by the recent progress of neural network models for biological tasks, we implement two neural network architectures that can treat such edit operations. The first proposal is the edit invariant neural networks, based on differentiable Needleman-Wunsch algorithms. The second is the use of deep CNNs with concatenations. Our analysis shows that CNNs can recognize star-free regular expressions, and that deeper CNNs can recognize more complex regular expressions including the insertion/deletion of characters. The experimental results for the protein secondary structure prediction task suggest the importance of insertion/deletion. The test accuracy on the widely-used CB513 dataset is 71.5%, which is 1.2-points better than the current best result on non-ensemble models.

Improved Expressivity Through Dendritic Neural Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #126

Xundong Wu · Xiangwen Liu · wei li · qing wu

A typical biological neuron, such as a pyramidal neuron of the neocortex, receives thousands of afferent synaptic inputs on its dendrite tree and sends the efferent axonal output downstream. In typical artificial neural networks, dendrite trees are modeled as linear structures that funnel weighted synaptic inputs to the cell bodies. However, numerous experimental and theoretical studies have shown that dendritic arbors are far more than simple linear accumulators. That is, synaptic inputs can actively modulate their neighboring synaptic activities; therefore, the dendritic structures are highly nonlinear. In this study, we model such local nonlinearity of dendritic trees with our dendritic neural network (DENN) structure and apply this structure to typical machine learning tasks. Equipped with localized nonlinearities, DENNs can attain greater model expressivity than regular neural networks while maintaining efficient network inference. Such strength is evidenced by the increased fitting power when we train DENNs with supervised machine learning tasks. We also empirically show that the locality structure can improve the generalization performance of DENNs, as exemplified by DENNs outranking naive deep neural network architectures when tested on 121 classification tasks from the UCI machine learning repository.

Neural Proximal Gradient Descent for Compressive Imaging

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #127

Morteza Mardani · Qingyun Sun · David Donoho · Vardan Papyan · Hatef Monajemi · Shreyas Vasanawala · John Pauly

Recovering high-resolution images from limited sensory data typically leads to a serious ill-posed inverse problem, demanding inversion algorithms that effectively capture the prior information. Learning a good inverse mapping from training data faces severe challenges, including: (i) scarcity of training data; (ii) need for plausible reconstructions that are physically feasible; (iii) need for fast reconstruction, especially in real-time applications. We develop a successful system solving all these challenges, using as basic architecture the repetitive application of alternating proximal and data fidelity constraints. We learn a proximal map that works well with real images based on residual networks with recurrent blocks. Extensive experiments are carried out under different settings: (a) reconstructing abdominal MRI of pediatric patients from highly undersampled k-space data and (b) super-resolving natural face images. Our key findings include: 1. a recurrent ResNet with a single residual block (10-fold repetition) yields an effective proximal which accurately reveals MR image details. 2. Our architecture significantly outperforms conventional non-recurrent deep ResNets by 2dB SNR; it is also trained much more rapidly. 3. It outperforms state-of-the-art compressed-sensing Wavelet-based methods by 4dB SNR, with 100x speedups in reconstruction time.

Sigsoftmax: Reanalysis of the Softmax Bottleneck

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #128

Sekitoshi Kanai · Yasuhiro Fujiwara · Yuki Yamanaka · Shuichi Adachi

Softmax is an output activation function for modeling categorical probability distributions in many applications of deep learning. However, a recent study revealed that softmax can be a bottleneck of representational capacity of neural networks in language modeling (the softmax bottleneck). In this paper, we propose an output activation function for breaking the softmax bottleneck without additional parameters. We re-analyze the softmax bottleneck from the perspective of the output set of log-softmax and identify the cause of the softmax bottleneck. On the basis of this analysis, we propose sigsoftmax, which is composed of a multiplication of an exponential function and sigmoid function. Sigsoftmax can break the softmax bottleneck. The experiments on language modeling demonstrate that sigsoftmax and mixture of sigsoftmax outperform softmax and mixture of softmax, respectively.

Visualizing the Loss Landscape of Neural Nets

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #129

Hao Li · Zheng Xu · Gavin Taylor · Christoph Studer · Tom Goldstein

Neural network training relies on our ability to find "good" minimizers of highly non-convex loss functions. It is well known that certain network architecture designs (e.g., skip connections) produce loss functions that train easier, and well-chosen training parameters (batch size, learning rate, optimizer) produce minimizers that generalize better. However, the reasons for these differences, and their effect on the underlying loss landscape, is not well understood. In this paper, we explore the structure of neural loss functions, and the effect of loss landscapes on generalization, using a range of visualization methods. First, we introduce a simple "filter normalization" method that helps us visualize loss function curvature, and make meaningful side-by-side comparisons between loss functions. Then, using a variety of visualizations, we explore how network architecture affects the loss landscape, and how training parameters affect the shape of minimizers.

Clebsch–Gordan Nets: a Fully Fourier Space Spherical Convolutional Neural Network

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #130

Risi Kondor · Zhen Lin · Shubhendu Trivedi

Recent work by Cohen et al. has achieved state-of-the-art results for learning spherical images in a rotation invariant way by using ideas from group representation theory and noncommutative harmonic analysis. In this paper we propose a generalization of this work that generally exhibits improved performace, but from an implementation point of view is actually simpler. An unusual feature of the proposed architecture is that it uses the Clebsch--Gordan transform as its only source of nonlinearity, thus avoiding repeated forward and backward Fourier transforms. The underlying ideas of the paper generalize to constructing neural networks that are invariant to the action of other compact groups.

Adaptive Sampling Towards Fast Graph Representation Learning

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #131

Wenbing Huang · Tong Zhang · Yu Rong · Junzhou Huang

Graph Convolutional Networks (GCNs) have become a crucial tool on learning representations of graph vertices. The main challenge of adapting GCNs on large-scale graphs is the scalability issue that it incurs heavy cost both in computation and memory due to the uncontrollable neighborhood expansion across layers. In this paper, we accelerate the training of GCNs through developing an adaptive layer-wise sampling method. By constructing the network layer by layer in a top-down passway, we sample the lower layer conditioned on the top one, where the sampled neighborhoods are shared by different parent nodes and the over expansion is avoided owing to the fixed-size sampling. More importantly, the proposed sampler is adaptive and applicable for explicit variance reduction, which in turn enhances the training of our method. Furthermore, we propose a novel and economical approach to promote the message passing over distant nodes by applying skip connections. Intensive experiments on several benchmarks verify the effectiveness of our method regarding the classification accuracy while enjoying faster convergence speed.

NAIS-Net: Stable Deep Networks from Non-Autonomous Differential Equations

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #132

Marco Ciccone · Marco Gallieri · Jonathan Masci · Christian Osendorfer · Faustino Gomez

This paper introduces Non-Autonomous Input-Output Stable Network (NAIS-Net), a very deep architecture where each stacked processing block is derived from a time-invariant non-autonomous dynamical system. Non-autonomy is implemented by skip connections from the block input to each of the unrolled processing stages and allows stability to be enforced so that blocks can be unrolled adaptively to a pattern-dependent processing depth. NAIS-Net induces non-trivial, Lipschitz input-output maps, even for an infinite unroll length. We prove that the network is globally asymptotically stable so that for every initial condition there is exactly one input-dependent equilibrium assuming tanh units, and multiple stable equilibria for ReL units. An efficient implementation that enforces the stability under derived conditions for both fully-connected and convolutional layers is also presented. Experimental results show how NAIS-Net exhibits stability in practice, yielding a significant reduction in generalization gap compared to ResNets.

Scaling provable adversarial defenses

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #133

Eric Wong · Frank Schmidt · Jan Hendrik Metzen · J. Zico Kolter

Recent work has developed methods for learning deep network classifiers that are \emph{provably} robust to norm-bounded adversarial perturbation; however, these methods are currently only possible for relatively small feedforward networks. In this paper, in an effort to scale these approaches to substantially larger models, we extend previous work in three main directly. First, we present a technique for extending these training procedures to much more general networks, with skip connections (such as ResNets) and general nonlinearities; the approach is fully modular, and can be implemented automatically analogously to automatic differentiation. Second, in the specific case of ℓ_∞ adversarial perturbations and networks with ReLU nonlinearities, we adopt a nonlinear random projection for training, which scales \emph{linearly} in the number of hidden units (previous approached scaled quadratically). Third, we show how to further improve robust error through cascade models. On both MNIST and CIFAR data sets, we train classifiers that improve substantially on the state of the art in provable robust adversarial error bounds: from 5.8% to 3.1% on MNIST (with ℓ_∞ perturbations of $\epsilon=0.1$), and from 80% to 36.4% on CIFAR (with ℓ_∞ perturbations of $\epsilon=2/255$).

Lipschitz regularity of deep neural networks: analysis and efficient estimation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #134

Aladin Virmaux · Kevin Scaman

Deep neural networks are notorious for being sensitive to small well-chosen perturbations, and estimating the regularity of such architectures is of utmost importance for safe and robust practical applications. In this paper, we investigate one of the key characteristics to assess the regularity of such methods: the Lipschitz constant of deep learning architectures. First, we show that, even for two layer neural networks, the exact computation of this quantity is NP-hard and state-of-art methods may significantly overestimate it. Then, we both extend and improve previous estimation methods by providing AutoLip, the first generic algorithm for upper bounding the Lipschitz constant of any automatically differentiable function. We provide a power method algorithm working with automatic differentiation, allowing efficient computations even on large convolutions. Second, for sequential neural networks, we propose an improved algorithm named SeqLip that takes advantage of the linear computation graph to split the computation per pair of consecutive layers. Third we propose heuristics on SeqLip in order to tackle very large networks. Our experiments show that SeqLip can significantly improve on the existing upper bounds. Finally, we provide an implementation of AutoLip in the PyTorch environment that may be used to better estimate the robustness of a given neural network to small perturbations or regularize it using more precise Lipschitz estimations. These results also hint at the difficulty to estimate the Lipschitz constant of deep networks.

Training DNNs with Hybrid Block Floating Point

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #135

Mario Drumond · Tao LIN · Martin Jaggi · Babak Falsafi

The wide adoption of DNNs has given birth to unrelenting computing requirements, forcing datacenter operators to adopt domain-specific accelerators to train them. These accelerators typically employ densely packed full-precision floating-point arithmetic to maximize performance per area. Ongoing research efforts seek to further increase that performance density by replacing floating-point with fixed-point arithmetic. However, a significant roadblock for these attempts has been fixed point's narrow dynamic range, which is insufficient for DNN training convergence. We identify block floating point (BFP) as a promising alternative representation since it exhibits wide dynamic range and enables the majority of DNN operations to be performed with fixed-point logic. Unfortunately, BFP alone introduces several limitations that preclude its direct applicability. In this work, we introduce HBFP, a hybrid BFP-FP approach, which performs all dot products in BFP and other operations in floating point. HBFP delivers the best of both worlds: the high accuracy of floating point at the superior hardware density of fixed point. For a wide variety of models, we show that HBFP matches floating point's accuracy while enabling hardware implementations that deliver up to 8.5x higher throughput.

Mesh-TensorFlow: Deep Learning for Supercomputers

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #136

Noam Shazeer · Youlong Cheng · Niki Parmar · Dustin Tran · Ashish Vaswani · Penporn Koanantakool · Peter Hawkins · HyoukJoong Lee · Mingsheng Hong · Cliff Young · Ryan Sepassi · Blake Hechtman

Batch-splitting (data-parallelism) is the dominant distributed Deep Neural Network (DNN) training strategy, due to its universal applicability and its amenability to Single-Program-Multiple-Data (SPMD) programming. However, batch-splitting suffers from problems including the inability to train very large models (due to memory constraints), high latency, and inefficiency at small batch sizes. All of these can be solved by more general distribution strategies (model-parallelism). Unfortunately, efficient model-parallel algorithms tend to be complicated to discover, describe, and to implement, particularly on large clusters. We introduce Mesh-TensorFlow, a language for specifying a general class of distributed tensor computations. Where data-parallelism can be viewed as splitting tensors and operations along the "batch" dimension, in Mesh-TensorFlow, the user can specify any tensor-dimensions to be split across any dimensions of a multi-dimensional mesh of processors. A Mesh-TensorFlow graph compiles into a SPMD program consisting of parallel operations coupled with collective communication primitives such as Allreduce. We use Mesh-TensorFlow to implement an efficient data-parallel, model-parallel version of the Transformer sequence-to-sequence model. Using TPU meshes of up to 512 cores, we train Transformer models with up to 5 billion parameters, surpassing SOTA results on WMT'14 English-to-French translation task and the one-billion-word Language modeling benchmark. Mesh-Tensorflow is available at https://github.com/tensorflow/mesh

---

Thwarting Adversarial Examples: An L_0-Robust Sparse Fourier Transform

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #137

Mitali Bafna · Jack Murtagh · Nikhil Vyas

We give a new algorithm for approximating the Discrete Fourier transform of an approximately sparse signal that is robust to worst-case L_0 corruptions, namely that some coordinates of the signal can be corrupt arbitrarily. Our techniques generalize to a wide range of linear transformations that are used in data analysis such as the Discrete Cosine and Sine transforms, the Hadamard transform, and their high-dimensional analogs. We use our algorithm to successfully defend against worst-case L_0 adversaries in the setting of image classification. We give experimental results on the Jacobian-based Saliency Map Attack (JSMA) and the CW L_0 attack on the MNIST and Fashion-MNIST datasets as well as the Adversarial Patch on the ImageNet dataset.

Bayesian Adversarial Learning

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #138

Nanyang Ye · Zhanxing Zhu

Deep neural networks have been known to be vulnerable to adversarial attacks, raising lots of security concerns in the practical deployment. Popular defensive approaches can be formulated as a (distributionally) robust optimization problem, which minimizes a ``point estimate'' of worst-case loss derived from either per-datum perturbation or adversary data-generating distribution within certain pre-defined constraints. This point estimate ignores potential test adversaries that are beyond the pre-defined constraints. The model robustness might deteriorate sharply in the scenario of stronger test adversarial data. In this work, a novel robust training framework is proposed to alleviate this issue, Bayesian Robust Learning, in which a distribution is put on the adversarial data-generating distribution to account for the uncertainty of the adversarial data-generating process.

The uncertainty directly helps to consider the potential adversaries that are stronger than the point estimate in the cases of distributionally robust optimization. The uncertainty of model parameters is also incorporated to accommodate the full Bayesian framework. We design a scalable Markov Chain Monte Carlo sampling strategy to obtain the posterior distribution over model parameters. Various experiments are conducted to verify the superiority of BAL over existing adversarial training methods. The code for BAL is available at \url{https://tinyurl.com/ycxsaewr }.

Dendritic cortical microcircuits approximate the backpropagation algorithm

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #139

João Sacramento · Rui Ponte Costa · Yoshua Bengio · Walter Senn

Deep learning has seen remarkable developments over the last years, many of them inspired by neuroscience. However, the main learning mechanism behind these advances – error backpropagation – appears to be at odds with neurobiology. Here, we introduce a multilayer neuronal network model with simplified dendritic compartments in which error-driven synaptic plasticity adapts the network towards a global desired output. In contrast to previous work our model does not require separate phases and synaptic learning is driven by local dendritic prediction errors continuously in time. Such errors originate at apical dendrites and occur due to a mismatch between predictive input from lateral interneurons and activity from actual top-down feedback. Through the use of simple dendritic compartments and different cell-types our model can represent both error and normal activity within a pyramidal neuron. We demonstrate the learning capabilities of the model in regression and classification tasks, and show analytically that it approximates the error backpropagation algorithm. Moreover, our framework is consistent with recent observations of learning between brain areas and the architecture of cortical microcircuits. Overall, we introduce a novel view of learning on dendritic cortical circuits and on how the brain may solve the long-standing synaptic credit assignment problem.

Learning a latent manifold of odor representations from neural responses in piriform cortex

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #140

Anqi Wu · Stan Pashkovski · Sandeep Datta · Jonathan W Pillow

A major difficulty in studying the neural mechanisms underlying olfactory perception is the lack of obvious structure in the relationship between odorants and the neural activity patterns they elicit. Here we use odor-evoked responses in piriform cortex to identify a latent manifold specifying latent distance relationships between olfactory stimuli. Our approach is based on the Gaussian process latent variable model, and seeks to map odorants to points in a low-dimensional embedding space, where distances between points in the embedding space relate to the similarity of population responses they elicit. The model is specified by an explicit continuous mapping from a latent embedding space to the space of high-dimensional neural population firing rates via nonlinear tuning curves, each parametrized by a Gaussian process. Population responses are then generated by the addition of correlated, odor-dependent Gaussian noise. We fit this model to large-scale calcium fluorescence imaging measurements of population activity in layers 2 and 3 of mouse piriform cortex following the presentation of a diverse set of odorants. The model identifies a low-dimensional embedding of each odor, and a smooth tuning curve over the latent embedding space that accurately captures each neuron's response to different odorants. The model captures both signal and noise correlations across more than 500 neurons. We validate the model using a cross-validation analysis known as co-smoothing to show that the model can accurately predict the responses of a population of held-out neurons to test odorants.

Size-Noise Tradeoffs in Generative Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #141

Bolton Bailey · Matus Telgarsky

This paper investigates the ability of generative networks to convert their input noise distributions into other distributions. Firstly, we demonstrate a construction that allows ReLU networks to increase the dimensionality of their noise distribution by implementing a ``space-filling'' function based on iterated tent maps. We show this construction is optimal by analyzing the number of affine pieces in functions computed by multivariate ReLU networks. Secondly, we provide efficient ways (using polylog$(1/\epsilon)$ nodes) for networks to pass between univariate uniform and normal distributions, using a Taylor series approximation and a binary search gadget for computing function inverses. Lastly, we indicate how high dimensional distributions can be efficiently transformed into low dimensional distributions.

On Neuronal Capacity

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #142

Pierre Baldi · Roman Vershynin

We define the capacity of a learning machine to be the logarithm of the number (or volume) of the functions it can implement. We review known results, and derive new results, estimating the capacity of several neuronal models: linear and polynomial threshold gates, linear and polynomial threshold gates with constrained weights (binary weights, positive weights), and ReLU neurons. We also derive capacity estimates and bounds for fully recurrent networks and layered feedforward networks.

Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #143

Yuanzhi Li · Yingyu Liang

Neural networks have many successful applications, while much less theoretical understanding has been gained. Towards bridging this gap, we study the problem of learning a two-layer overparameterized ReLU neural network for multi-class classification via stochastic gradient descent (SGD) from random initialization. In the overparameterized setting, when the data comes from mixtures of well-separated distributions, we prove that SGD learns a network with a small generalization error, albeit the network has enough capacity to fit arbitrary labels. Furthermore, the analysis provides interesting insights into several aspects of learning neural networks and can be verified based on empirical studies on synthetic data and on the MNIST dataset.

Deep, complex, invertible networks for inversion of transmission effects in multimode optical fibres

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #144

Oisín Moran · Piergiorgio Caramazza · Daniele Faccio · Roderick Murray-Smith

We use complex-weighted, deep networks to invert the effects of multimode optical fibre distortion of a coherent input image. We generated experimental data based on collections of optical fibre responses to greyscale input images generated with coherent light, by measuring only image amplitude (not amplitude and phase as is typical) at the output of \SI{1}{\metre} and \SI{10}{\metre} long, \SI{105}{\micro\metre} diameter multimode fibre. This data is made available as the {\it Optical fibre inverse problem} Benchmark collection. The experimental data is used to train complex-weighted models with a range of regularisation approaches. A {\it unitary regularisation} approach for complex-weighted networks is proposed which performs well in robustly inverting the fibre transmission matrix, which fits well with the physical theory. A key benefit of the unitary constraint is that it allows us to learn a forward unitary model and analytically invert it to solve the inverse problem. We demonstrate this approach, and show how it can improve performance by incorporating knowledge of the phase shift induced by the spatial light modulator.

Learning towards Minimum Hyperspherical Energy

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #145

Weiyang Liu · Rongmei Lin · Zhen Liu · Lixin Liu · Zhiding Yu · Bo Dai · Le Song

Neural networks are a powerful class of nonlinear functions that can be trained end-to-end on various applications. While the over-parametrization nature in many neural networks renders the ability to fit complex functions and the strong representation power to handle challenging tasks, it also leads to highly correlated neurons that can hurt the generalization ability and incur unnecessary computation cost. As a result, how to regularize the network to avoid undesired representation redundancy becomes an important issue. To this end, we draw inspiration from a well-known problem in physics -- Thomson problem, where one seeks to find a state that distributes N electrons on a unit sphere as evenly as possible with minimum potential energy. In light of this intuition, we reduce the redundancy regularization problem to generic energy minimization, and propose a minimum hyperspherical energy (MHE) objective as generic regularization for neural networks. We also propose a few novel variants of MHE, and provide some insights from a theoretical point of view. Finally, we apply neural networks with MHE regularization to several challenging tasks. Extensive experiments demonstrate the effectiveness of our intuition, by showing the superior performance with MHE regularization.

Soft-Gated Warping-GAN for Pose-Guided Person Image Synthesis

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #146

Haoye Dong · Xiaodan Liang · Ke Gong · Hanjiang Lai · Jia Zhu · Jian Yin

Despite remarkable advances in image synthesis research, existing works often fail in manipulating images under the context of large geometric transformations. Synthesizing person images conditioned on arbitrary poses is one of the most representative examples where the generation quality largely relies on the capability of identifying and modeling arbitrary transformations on different body parts. Current generative models are often built on local convolutions and overlook the key challenges (e.g. heavy occlusions, different views or dramatic appearance changes) when distinct geometric changes happen for each part, caused by arbitrary pose manipulations. This paper aims to resolve these challenges induced by geometric variability and spatial displacements via a new Soft-Gated Warping Generative Adversarial Network (Warping-GAN), which is composed of two stages: 1) it first synthesizes a target part segmentation map given a target pose, which depicts the region-level spatial layouts for guiding image synthesis with higher-level structure constraints; 2) the Warping-GAN equipped with a soft-gated warping-block learns feature-level mapping to render textures from the original image into the generated segmentation map. Warping-GAN is capable of controlling different transformation degrees given distinct target poses. Moreover, the proposed warping-block is light-weight and flexible enough to be injected into any networks. Human perceptual studies and quantitative evaluations demonstrate the superiority of our Warping-GAN that significantly outperforms all existing methods on two large datasets.

Deep Attentive Tracking via Reciprocative Learning

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #147

Shi Pu · Yibing Song · Chao Ma · Honggang Zhang · Ming-Hsuan Yang

Visual attention, derived from cognitive neuroscience, facilitates human perception on the most pertinent subset of the sensory data. Recently, significant efforts have been made to exploit attention schemes to advance computer vision systems. For visual tracking, it is often challenging to track target objects undergoing large appearance changes. Attention maps facilitate visual tracking by selectively paying attention to temporal robust features. Existing tracking-by-detection approaches mainly use additional attention modules to generate feature weights as the classifiers are not equipped with such mechanisms. In this paper, we propose a reciprocative learning algorithm to exploit visual attention for training deep classifiers. The proposed algorithm consists of feed-forward and backward operations to generate attention maps, which serve as regularization terms coupled with the original classification loss function for training. The deep classifier learns to attend to the regions of target objects robust to appearance changes. Extensive experiments on large-scale benchmark datasets show that the proposed attentive tracking method performs favorably against the state-of-the-art approaches.

Robot Learning in Homes: Improving Generalization and Reducing Dataset Bias

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #148

Abhinav Gupta · Adithyavairavan Murali · Dhiraj Prakashchand Gandhi · Lerrel Pinto

Data-driven approaches to solving robotic tasks have gained a lot of traction in recent years. However, most existing policies are trained on large-scale datasets collected in curated lab settings. If we aim to deploy these models in unstructured visual environments like people's homes, they will be unable to cope with the mismatch in data distribution. In such light, we present the first systematic effort in collecting a large dataset for robotic grasping in homes. First, to scale and parallelize data collection, we built a low cost mobile manipulator assembled for under 3K USD. Second, data collected using low cost robots suffer from noisy labels due to imperfect execution and calibration errors. To handle this, we develop a framework which factors out the noise as a latent variable. Our model is trained on 28K grasps collected in several houses under an array of different environmental conditions. We evaluate our models by physically executing grasps on a collection of novel objects in multiple unseen homes. The models trained with our home dataset showed a marked improvement of 43.7% over a baseline model trained with data collected in lab. Our architecture which explicitly models the latent noise in the dataset also performed 10% better than one that did not factor out the noise. We hope this effort inspires the robotics community to look outside the lab and embrace learning based approaches to handle inaccurate cheap robots.

A flexible model for training action localization with varying levels of supervision

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #149

Guilhem Chéron · Jean-Baptiste Alayrac · Ivan Laptev · Cordelia Schmid

Spatio-temporal action detection in videos is typically addressed in a fully-supervised setup with manual annotation of training videos required at every frame. Since such annotation is extremely tedious and prohibits scalability, there is a clear need to minimize the amount of manual supervision. In this work we propose a unifying framework that can handle and combine varying types of less demanding weak supervision. Our model is based on discriminative clustering and integrates different types of supervision as constraints on the optimization. We investigate applications of such a model to training setups with alternative supervisory signals ranging from video-level class labels over temporal points or sparse action bounding boxes to the full per-frame annotation of action bounding boxes. Experiments on the challenging UCF101-24 and DALY datasets demonstrate competitive performance of our method at a fraction of supervision used by previous methods. The flexibility of our model enables joint learning from data with different levels of annotation. Experimental results demonstrate a significant gain by adding a few fully supervised examples to otherwise weakly labeled videos.

Bilinear Attention Networks

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #150

Jin-Hwa Kim · Jaehyun Jun · Byoung-Tak Zhang

Attention networks in multimodal learning provide an efficient way to utilize given visual information selectively. However, the computational cost to learn attention distributions for every pair of multimodal input channels is prohibitively expensive. To solve this problem, co-attention builds two separate attention distributions for each modality neglecting the interaction between multimodal inputs. In this paper, we propose bilinear attention networks (BAN) that find bilinear attention distributions to utilize given vision-language information seamlessly. BAN considers bilinear interactions among two groups of input channels, while low-rank bilinear pooling extracts the joint representations for each pair of channels. Furthermore, we propose a variant of multimodal residual networks to exploit eight-attention maps of the BAN efficiently. We quantitatively and qualitatively evaluate our model on visual question answering (VQA 2.0) and Flickr30k Entities datasets, showing that BAN significantly outperforms previous methods and achieves new state-of-the-arts on both datasets.

MacNet: Transferring Knowledge from Machine Comprehension to Sequence-to-Sequence Models

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #151

Boyuan Pan · Yazheng Yang · Hao Li · Zhou Zhao · Yueting Zhuang · Deng Cai · Xiaofei He

Machine Comprehension (MC) is one of the core problems in natural language processing, requiring both understanding of the natural language and knowledge about the world. Rapid progress has been made since the release of several benchmark datasets, and recently the state-of-the-art models even surpass human performance on the well-known SQuAD evaluation. In this paper, we transfer knowledge learned from machine comprehension to the sequence-to-sequence tasks to deepen the understanding of the text. We propose MacNet: a novel encoder-decoder supplementary architecture to the widely used attention-based sequence-to-sequence models. Experiments on neural machine translation (NMT) and abstractive text summarization show that our proposed framework can significantly improve the performance of the baseline models, and our method for the abstractive text summarization achieves the state-of-the-art results on the Gigaword dataset.

Diffusion Maps for Textual Network Embedding

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #152

Xinyuan Zhang · Yitong Li · Dinghan Shen · Lawrence Carin

Textual network embedding leverages rich text information associated with the network to learn low-dimensional vectorial representations of vertices. Rather than using typical natural language processing (NLP) approaches, recent research exploits the relationship of texts on the same edge to graphically embed text. However, these models neglect to measure the complete level of connectivity between any two texts in the graph. We present diffusion maps for textual network embedding (DMTE), integrating global structural information of the graph to capture the semantic relatedness between texts, with a diffusion-convolution operation applied on the text inputs. In addition, a new objective function is designed to efficiently preserve the high-order proximity using the graph diffusion. Experimental results show that the proposed approach outperforms state-of-the-art methods on the vertex-classification and link-prediction tasks.

FRAGE: Frequency-Agnostic Word Representation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #153

Chengyue Gong · Di He · Xu Tan · Tao Qin · Liwei Wang · Tie-Yan Liu

Continuous word representation (aka word embedding) is a basic building block in many neural network-based models used in natural language processing tasks. Although it is widely accepted that words with similar semantics should be close to each other in the embedding space, we find that word embeddings learned in several tasks are biased towards word frequency: the embeddings of high-frequency and low-frequency words lie in different subregions of the embedding space, and the embedding of a rare word and a popular word can be far from each other even if they are semantically similar. This makes learned word embeddings ineffective, especially for rare words, and consequently limits the performance of these neural network models. In order to mitigate the issue, in this paper, we propose a neat, simple yet effective adversarial training method to blur the boundary between the embeddings of high-frequency words and low-frequency words. We conducted comprehensive studies on ten datasets across four natural language processing tasks, including word similarity, language modeling, machine translation and text classification. Results show that we achieve higher performance than the baselines in all tasks.

A Retrieve-and-Edit Framework for Predicting Structured Outputs

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #154

Tatsunori Hashimoto · Kelvin Guu · Yonatan Oren · Percy Liang

For the task of generating complex outputs such as source code, editing existing outputs can be easier than generating complex outputs from scratch. With this motivation, we propose an approach that first retrieves a training example based on the input (e.g., natural language description) and then edits it to the desired output (e.g., code). Our contribution is a computationally efficient method for learning a retrieval model that embeds the input in a task-dependent way without relying on a hand-crafted metric or incurring the expense of jointly training the retriever with the editor. Our retrieve-and-edit framework can be applied on top of any base model. We show that on a new autocomplete task for GitHub Python code and the Hearthstone cards benchmark, retrieve-and-edit significantly boosts the performance of a vanilla sequence-to-sequence model on both tasks.

Unsupervised Text Style Transfer using Language Models as Discriminators

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #155

Zichao Yang · Zhiting Hu · Chris Dyer · Eric Xing · Taylor Berg-Kirkpatrick

Binary classifiers are employed as discriminators in GAN-based unsupervised style transfer models to ensure that transferred sentences are similar to sentences in the target domain. One difficulty with the binary discriminator is that error signal is sometimes insufficient to train the model to produce rich-structured language. In this paper, we propose a technique of using a target domain language model as the discriminator to provide richer, token-level feedback during the learning process. Because our language model scores sentences directly using a product of locally normalized probabilities, it offers more stable and more useful training signal to the generator. We train the generator to minimize the negative log likelihood (NLL) of generated sentences evaluated by a language model. By using continuous approximation of the discrete samples, our model can be trained using back-propagation in an end-to-end way. Moreover, we find empirically with a language model as a structured discriminator, it is possible to eliminate the adversarial training steps using negative samples, thus making training more stable. We compare our model with previous work using convolutional neural networks (CNNs) as discriminators and show our model outperforms them significantly in three tasks including word substitution decipherment, sentiment modification and related language translation.

Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #156

Yu-An Chung · Wei-Hung Weng · Schrasing Tong · James Glass

Recent research has shown that word embedding spaces learned from text corpora of different languages can be aligned without any parallel data supervision. Inspired by the success in unsupervised cross-lingual word embeddings, in this paper we target learning a cross-modal alignment between the embedding spaces of speech and text learned from corpora of their respective modalities in an unsupervised fashion. The proposed framework learns the individual speech and text embedding spaces, and attempts to align the two spaces via adversarial training, followed by a refinement procedure. We show how our framework could be used to perform the tasks of spoken word classification and translation, and the experimental results on these two tasks demonstrate that the performance of our unsupervised alignment approach is comparable to its supervised counterpart. Our framework is especially useful for developing automatic speech recognition (ASR) and speech-to-text translation systems for low- or zero-resource languages, which have little parallel audio-text data for training modern supervised ASR and speech-to-text translation models, but account for the majority of the languages spoken across the world.

GradiVeQ: Vector Quantization for Bandwidth-Efficient Gradient Aggregation in Distributed CNN Training

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #157

Mingchao Yu · Zhifeng Lin · Krishna Narra · Songze Li · Youjie Li · Nam Sung Kim · Alexander Schwing · Murali Annavaram · Salman Avestimehr

Data parallelism can boost the training speed of convolutional neural networks (CNN), but could suffer from significant communication costs caused by gradient aggregation. To alleviate this problem, several scalar quantization techniques have been developed to compress the gradients. But these techniques could perform poorly when used together with decentralized aggregation protocols like ring all-reduce (RAR), mainly due to their inability to directly aggregate compressed gradients. In this paper, we empirically demonstrate the strong linear correlations between CNN gradients, and propose a gradient vector quantization technique, named GradiVeQ, to exploit these correlations through principal component analysis (PCA) for substantial gradient dimension reduction. GradiveQ enables direct aggregation of compressed gradients, hence allows us to build a distributed learning system that parallelizes GradiveQ gradient compression and RAR communications. Extensive experiments on popular CNNs demonstrate that applying GradiveQ slashes the wall-clock gradient aggregation time of the original RAR by more than 5x without noticeable accuracy loss, and reduce the end-to-end training time by almost 50%. The results also show that \GradiveQ is compatible with scalar quantization techniques such as QSGD (Quantized SGD), and achieves a much higher speed-up gain under the same compression ratio.

Gradient Sparsification for Communication-Efficient Distributed Optimization

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #158

Jianqiao Wangni · Jialei Wang · Ji Liu · Tong Zhang

Modern large-scale machine learning applications require stochastic optimization algorithms to be implemented on distributed computational architectures. A key bottleneck is the communication overhead for exchanging information such as stochastic gradients among different workers. In this paper, to reduce the communication cost, we propose a convex optimization formulation to minimize the coding length of stochastic gradients. The key idea is to randomly drop out coordinates of the stochastic gradient vectors and amplify the remaining coordinates appropriately to ensure the sparsified gradient to be unbiased. To solve the optimal sparsification efficiently, several simple and fast algorithms are proposed for an approximate solution, with a theoretical guarantee for sparseness. Experiments on ℓ_2 regularized logistic regression, support vector machines, and convolutional neural networks validate our sparsification approaches.

Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #159

Ehsan Hajiramezanali · Siamak Zamani Dadaneh · Alireza Karbalayghareh · Mingyuan Zhou · Xiaoning Qian

Precision medicine aims for personalized prognosis and therapeutics by utilizing recent genome-scale high-throughput profiling techniques, including next-generation sequencing (NGS). However, translating NGS data faces several challenges. First, NGS count data are often overdispersed, requiring appropriate modeling. Second, compared to the number of involved molecules and system complexity, the number of available samples for studying complex disease, such as cancer, is often limited, especially considering disease heterogeneity. The key question is whether we may integrate available data from all different sources or domains to achieve reproducible disease prognosis based on NGS count data. In this paper, we develop a Bayesian Multi-Domain Learning (BMDL) model that derives domain-dependent latent representations of overdispersed count data based on hierarchical negative binomial factorization for accurate cancer subtyping even if the number of samples for a specific cancer type is small. Experimental results from both our simulated and NGS datasets from The Cancer Genome Atlas (TCGA) demonstrate the promising potential of BMDL for effective multi-domain learning without ``negative transfer'' effects often seen in existing multi-task learning and transfer learning methods.

Parsimonious Quantile Regression of Financial Asset Tail Dynamics via Sequential Learning

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #160

Xing Yan · Weizhong Zhang · Lin Ma · Wei Liu · Qi Wu

We propose a parsimonious quantile regression framework to learn the dynamic tail behaviors of financial asset returns. Our model captures well both the time-varying characteristic and the asymmetrical heavy-tail property of financial time series. It combines the merits of a popular sequential neural network model, i.e., LSTM, with a novel parametric quantile function that we construct to represent the conditional distribution of asset returns. Our model also captures individually the serial dependences of higher moments, rather than just the volatility. Across a wide range of asset classes, the out-of-sample forecasts of conditional quantiles or VaR of our model outperform the GARCH family. Further, the proposed approach does not suffer from the issue of quantile crossing, nor does it expose to the ill-posedness comparing to the parametric probability density function approach.

Global Geometry of Multichannel Sparse Blind Deconvolution on the Sphere

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #161

Yanjun Li · Yoram Bresler

Multichannel blind deconvolution is the problem of recovering an unknown signal f and multiple unknown channels x_i from convolutional measurements $y_i=x_i \circledast f$ ($i=1,2,\dots,N$). We consider the case where the x_i's are sparse, and convolution with f is invertible. Our nonconvex optimization formulation solves for a filter h on the unit sphere that produces sparse output $y_i\circledast h$. Under some technical assumptions, we show that all local minima of the objective function correspond to the inverse filter of f up to an inherent sign and shift ambiguity, and all saddle points have strictly negative curvatures. This geometric structure allows successful recovery of f and x_i using a simple manifold gradient descent algorithm with random initialization. Our theoretical findings are complemented by numerical experiments, which demonstrate superior performance of the proposed approach over the previous methods.

Phase Retrieval Under a Generative Prior

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #162

Paul Hand · Oscar Leong · Vlad Voroninski

We introduce a novel deep-learning inspired formulation of the \textit{phase retrieval problem}, which asks to recover a signal $y_0 \in \R^n$ from m quadratic observations, under structural assumptions on the underlying signal. As is common in many imaging problems, previous methodologies have considered natural signals as being sparse with respect to a known basis, resulting in the decision to enforce a generic sparsity prior. However, these methods for phase retrieval have encountered possibly fundamental limitations, as no computationally efficient algorithm for sparse phase retrieval has been proven to succeed with fewer than $O(k^2\log n)$ generic measurements, which is larger than the theoretical optimum of $O(k \log n)$. In this paper, we sidestep this issue by considering a prior that a natural signal is in the range of a generative neural network $G : \R^k \rightarrow \R^n$. We introduce an empirical risk formulation that has favorable global geometry for gradient methods, as soon as $m = O(k)$, under the model of a multilayer fully-connected neural network with random weights. Specifically, we show that there exists a descent direction outside of a small neighborhood around the true k-dimensional latent code and a negative multiple thereof. This formulation for structured phase retrieval thus benefits

from two effects: generative priors can more tightly represent natural signals than sparsity priors, and this empirical risk formulation can exploit those generative priors at an information theoretically optimal sample complexity, unlike for a sparsity prior. We corroborate these results with experiments showing that exploiting generative models in phase retrieval tasks outperforms both sparse and general phase retrieval methods.

Theoretical Linear Convergence of Unfolded ISTA and Its Practical Weights and Thresholds

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #163

Xiaohan Chen · Jialin Liu · Zhangyang Wang · Wotao Yin

In recent years, unfolding iterative algorithms as neural networks has become an empirical success in solving sparse recovery problems. However, its theoretical understanding is still immature, which prevents us from fully utilizing the power of neural networks. In this work, we study unfolded ISTA (Iterative Shrinkage Thresholding Algorithm) for sparse signal recovery. We introduce a weight structure that is necessary for asymptotic convergence to the true sparse signal. With this structure, unfolded ISTA can attain a linear convergence, which is better than the sublinear convergence of ISTA/FISTA in general cases. Furthermore, we propose to incorporate thresholding in the network to perform support selection, which is easy to implement and able to boost the convergence rate both theoretically and empirically. Extensive simulations, including sparse vector recovery and a compressive sensing experiment on real image data, corroborate our theoretical results and demonstrate their practical usefulness. We have made our codes publicly available: https://github.com/xchen-tamu/linear-lista-cpss.

Modern Neural Networks Generalize on Small Data Sets

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #164

Matthew Olson · Abraham Wyner · Richard Berk

In this paper, we use a linear program to empirically decompose fitted neural networks into ensembles of low-bias sub-networks. We show that these sub-networks are relatively uncorrelated which leads to an internal regularization process, very much like a random forest, which can explain why a neural network is surprisingly resistant to overfitting. We then demonstrate this in practice by applying large neural networks, with hundreds of parameters per training observation, to a collection of 116 real-world data sets from the UCI Machine Learning Repository. This collection of data sets contains a much smaller number of training examples than the types of image classification tasks generally studied in the deep learning literature, as well as non-trivial label noise. We show that even in this setting deep neural nets are capable of achieving superior classification accuracy

without overfitting.

Co-regularized Alignment for Unsupervised Domain Adaptation

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #165

Abhishek Kumar · Prasanna Sattigeri · Kahini Wadhawan · Leonid Karlinsky · Rogerio Feris · Bill Freeman · Gregory Wornell

Deep neural networks, trained with large amount of labeled data, can fail to generalize well when tested with examples from a target domain whose distribution differs from the training data distribution, referred as the source domain. It can be expensive or even infeasible to obtain required amount of labeled data in all possible domains. Unsupervised domain adaptation sets out to address this problem, aiming to learn a good predictive model for the target domain using labeled examples from the source domain but only unlabeled examples from the target domain. Domain alignment approaches this problem by matching the source and target feature distributions, and has been used as a key component in many state-of-the-art domain adaptation methods. However, matching the marginal feature distributions does not guarantee that the corresponding class conditional distributions will be aligned across the two domains. We propose co-regularized domain alignment for unsupervised domain adaptation, which constructs multiple diverse feature spaces and aligns source and target distributions in each of them individually, while encouraging that alignments agree with each other with regard to the class predictions on the unlabeled target examples. The proposed method is generic and can be used to improve any domain adaptation method which uses domain alignment. We instantiate it in the context of a recent state-of-the-art method and observe that it provides significant performance improvements on several domain adaptation benchmarks.

Neural Networks Trained to Solve Differential Equations Learn General Representations

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #166

Martin Magill · Faisal Qureshi · Hendrick de Haan

We introduce a technique based on the singular vector canonical correlation analysis (SVCCA) for measuring the generality of neural network layers across a continuously-parametrized set of tasks. We illustrate this method by studying generality in neural networks trained to solve parametrized boundary value problems based on the Poisson partial differential equation. We find that the first hidden layers are general, and that they learn generalized coordinates over the input domain. Deeper layers are successively more specific. Next, we validate our method against an existing technique that measures layer generality using transfer learning experiments. We find excellent agreement between the two methods, and note that our method is much faster, particularly for continuously-parametrized problems. Finally, we also apply our method to networks trained on MNIST, and show it is consistent with, and complimentary to, another study of intrinsic dimensionality.

Spectral Filtering for General Linear Dynamical Systems

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #167

Elad Hazan · HOLDEN LEE · Karan Singh · Cyril Zhang · Yi Zhang

We give a polynomial-time algorithm for learning latent-state linear dynamical systems without system identification, and without assumptions on the spectral radius of the system's transition matrix. The algorithm extends the recently introduced technique of spectral filtering, previously applied only to systems with a symmetric transition matrix, using a novel convex relaxation to allow for the efficient identification of phases.

Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behavior

Poster | Tue Dec 4th 10:45 AM -- 12:45 PM @ Room 210 & 230 AB #168

Sid Reddy · Anca Dragan · Sergey Levine

Inferring intent from observed behavior has been studied extensively within the frameworks of Bayesian inverse planning and inverse reinforcement learning. These methods infer a goal or reward function that best explains the actions of the observed agent, typically a human demonstrator. Another agent can use this inferred intent to predict, imitate, or assist the human user. However, a central assumption in inverse reinforcement learning is that the demonstrator is close to optimal. While models of suboptimal behavior exist, they typically assume that suboptimal actions are the result of some type of random noise or a known cognitive bias, like temporal inconsistency. In this paper, we take an alternative approach, and model suboptimal behavior as the result of internal model misspecification: the reason that user actions might deviate from near-optimal actions is that the user has an incorrect set of beliefs about the rules -- the dynamics -- governing how actions affect the environment. Our insight is that while demonstrated actions may be suboptimal in the real world, they may actually be near-optimal with respect to the user's internal model of the dynamics. By estimating these internal beliefs from observed behavior, we arrive at a new method for inferring intent. We demonstrate in simulation and in a user study with 12 participants that this approach enables us to more accurately model human intent, and can be used in a variety of applications, including offering assistance in a shared autonomy framework and inferring human preferences.

Lunch on your own

Break | Tue Dec 4th 12:45 -- 02:15 PM @

—

Lunch on your own

Break | Tue Dec 4th 12:45 -- 02:15 PM @

—

Town Hall

Break | Tue Dec 4th 01:15 -- 02:15 PM @ Room 517 CD

—

What Bodies Think About: Bioelectric Computation Outside the Nervous System, Primitive Cognition, and Synthetic Morphology

Invited Talk | Tue Dec 4th 02:15 -- 03:05 PM @ Rooms 220 CDE

Michael Levin

Brains are not unique in their computational abilities. Bacteria, plants, and unicellular organisms exhibit learning and plasticity; nervous systems speed-optimized information-processing that is ubiquitous across the tree of life and was already occurring at multiple scales before neurons evolved. Non-neural computation is especially critical for enabling individual cells to coordinate their activity toward the creation and repair of complex large-scale anatomies. We have found that bioelectric signaling enables all types of cells to form networks that store pattern memories that guide large-scale growth and form. In this talk, I will introduce the basics of developmental bioelectricity, and show how novel conceptual and methodological advances have enabled rewriting pattern memories that guide morphogenesis without genomic editing. In effect, these strategies allow reprogramming the bioelectric software that implements multicellular patterning goal states. I will show examples of applications in regenerative medicine and cognitive neuroplasticity, and illustrate future impacts on synthetic bioengineering, robotics, and machine learning.

Coffee Break

Break | Tue Dec 4th 03:05 -- 03:30 PM @

—

Neural Voice Cloning with a Few Samples

Spotlight | Tue Dec 4th 03:30 -- 03:35 PM @ Room 220 CD

Sercan Arik · Jitong Chen · Kainan Peng · Wei Ping · Yanqi Zhou

Voice cloning is a highly desired feature for personalized speech interfaces. We introduce a neural voice cloning system that learns to synthesize a person's voice from only a few audio samples. We study two approaches: speaker adaptation and speaker encoding. Speaker adaptation is based on fine-tuning a multi-speaker generative model. Speaker encoding is based on training a separate model to directly infer a new speaker embedding, which will be applied to a multi-speaker generative model. In terms of naturalness of the speech and similarity to the original speaker, both approaches can achieve good performance, even with a few cloning audios. While speaker adaptation can achieve slightly better naturalness and similarity, cloning time and required memory for the speaker encoding approach are significantly less, making it more favorable for low-resource deployment.

Evolved Policy Gradients

Spotlight | Tue Dec 4th 03:30 -- 03:35 PM @ Room 220 E

Rein Houthooft · Yuhua Chen · Phillip Isola · Bradly Stadie · Filip Wolski · OpenAI Jonathan Ho · Pieter Abbeel

We propose a metalearning approach for learning gradient-based reinforcement learning (RL) algorithms. The idea is to evolve a differentiable loss function, such that an agent, which optimizes its policy to minimize this loss, will achieve high rewards. The loss is parametrized via temporal convolutions over the agent's experience. Because this loss is highly flexible in its ability to take into account the agent's history, it enables fast task learning. Empirical results show that our evolved policy gradient algorithm (EPG) achieves faster learning on several randomized environments compared to an off-the-shelf policy gradient method. We also demonstrate that EPG's learned loss can generalize to out-of-distribution test time tasks, and exhibits qualitatively different behavior from other popular metalearning algorithms.

Differentially Private Testing of Identity and Closeness of Discrete Distributions

Spotlight | Tue Dec 4th 03:30 -- 03:35 PM @ Room 517 CD

Jayadev Acharya · Ziteng Sun · Huanyu Zhang

We study the fundamental problems of identity testing (goodness of fit), and closeness testing (two sample test) of distributions over k elements, under differential privacy. While the problems have a long history in statistics, finite sample bounds for these problems have only been established recently.

In this work, we derive upper and lower bounds on the sample complexity of both the problems under (ε, δ)-differential privacy. We provide optimal sample complexity algorithms for identity testing problem for all parameter ranges, and the first results for closeness testing. Our closeness testing bounds are optimal in the sparse regime where the number of samples is at most k.

Our upper bounds are obtained by privatizing non-private estimators for these problems. The non-private estimators are chosen to have small sensitivity. We propose a general framework to establish lower bounds on the sample complexity of statistical tasks under differential privacy. We show a bound on differentially private algorithms in terms of a coupling between the two hypothesis classes we aim to test. By constructing carefully chosen priors over the hypothesis classes, and using Le Cam's two point theorem we provide a general mechanism for proving lower bounds. We believe that the framework can be used to obtain strong lower bounds for other statistical tasks under privacy.

Answerer in Questioner's Mind: Information Theoretic Approach to Goal-Oriented Visual Dialog

Spotlight | Tue Dec 4th 03:35 -- 03:40 PM @ Room 220 CD

Sang-Woo Lee · Yu-Jung Heo · Byoung-Tak Zhang

Goal-oriented dialog has been given attention due to its numerous applications in artificial intelligence. Goal-oriented dialogue tasks occur when a questioner asks an action-oriented question and an answerer responds with the intent of letting the questioner know a correct action to take. To ask the adequate question, deep learning and reinforcement learning have been recently applied. However, these approaches struggle to find a competent recurrent neural questioner, owing to the complexity of learning a series of sentences. Motivated by theory of mind, we propose "Answerer in Questioner's Mind" (AQM), a novel information theoretic algorithm for goal-oriented dialog. With

AQM, a questioner asks and infers based on an approximated probabilistic model of the answerer. The questioner figures out the answerer's intention via selecting a plausible question by explicitly calculating the information gain of the candidate intentions and possible answers to each question. We test our framework on two goal-oriented visual dialog tasks: "MNIST Counting Dialog" and "GuessWhat?!". In our experiments, AQM outperforms comparative algorithms by a large margin.

Adapted Deep Embeddings: A Synthesis of Methods for k-Shot Inductive Transfer Learning

Spotlight | Tue Dec 4th 03:35 -- 03:40 PM @ Room 220 E

Tyler Scott · Karl Ridgeway · Michael Mozer

The focus in machine learning has branched beyond training classifiers on a single task to investigating how previously acquired knowledge in a source domain can be leveraged to facilitate learning in a related target domain, known as inductive transfer learning. Three active lines of research have independently explored transfer learning using neural networks. In weight transfer, a model trained on the source domain is used as an initialization point for a network to be trained on the target domain. In deep metric learning, the source domain is used to construct an embedding that captures class structure in both the source and target domains. In few-shot learning, the focus is on generalizing well in the target domain based on a limited number of labeled examples. We compare state-of-the-art methods from these three paradigms and also explore hybrid adapted-embedding methods that use limited target-domain data to fine tune embeddings constructed from source-domain data. We conduct a systematic comparison of methods in a variety of domains, varying the number of labeled instances available in the target domain (k), as well as the number of target-domain classes. We reach three principal conclusions: (1) Deep embeddings are far superior, compared to weight transfer, as a starting point for inter-domain transfer or model re-use (2) Our hybrid methods robustly outperform every few-shot learning and every deep metric learning method previously proposed, with a mean error reduction of 34% over state-of-the-art. (3) Among loss functions for discovering embeddings, the histogram loss (Ustinova & Lempitsky, 2016) is most robust. We hope our results will motivate a unification of research in weight transfer, deep metric learning, and few-shot learning.

Local Differential Privacy for Evolving Data

Spotlight | Tue Dec 4th 03:35 -- 03:40 PM @ Room 517 CD

Matthew Joseph · Aaron Roth · Jonathan Ullman · Bo Waggoner

There are now several large scale deployments of differential privacy used to collect statistical information about users. However, these deployments periodically recollect the data and recompute the statistics using algorithms designed for a single use. As a result, these systems do not provide meaningful privacy guarantees over long time scales. Moreover, existing techniques to mitigate this effect do not apply in the ``local model'' of differential privacy that these systems use. In this paper, we introduce a new technique for local differential privacy that makes it possible to maintain up-to-date statistics over time, with privacy guarantees that degrade only in the number of changes in the underlying distribution rather than the number of collection periods. We use our technique for tracking a changing statistic in the setting where users are partitioned into an unknown collection of groups, and at every time period each user draws a single bit from a common (but changing) group-specific distribution. We also provide an application to frequency and heavy-hitter estimation.

Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding

Spotlight | Tue Dec 4th 03:40 -- 03:45 PM @ Room 220 CD

Kexin Yi · Jiajun Wu · Chuang Gan · Antonio Torralba · Pushmeet Kohli · Josh Tenenbaum

We marry two powerful ideas: deep representation learning for visual recognition and language understanding, and symbolic program execution for reasoning. Our neural-symbolic visual question answering (NS-VQA) system first recovers a structural scene representation from the image and a program trace from the question. It then executes the program on the scene representation to obtain an answer. Incorporating symbolic structure as prior knowledge offers three unique advantages. First, executing programs on a symbolic space is more robust to long program traces; our model can solve complex reasoning tasks better, achieving an accuracy of 99.8% on the CLEVR dataset. Second, the model is more data- and memory-efficient: it performs well after learning on a small number of training data; it can also encode an image into a compact representation, requiring less storage than existing methods for offline question answering. Third, symbolic program execution offers full transparency to the reasoning process; we are thus able to interpret and diagnose each execution step.

Bayesian Model-Agnostic Meta-Learning

Spotlight | Tue Dec 4th 03:40 -- 03:45 PM @ Room 220 E

Jaesik Yoon · Taesup Kim · Ousmane Dia · Sungwoong Kim · Yoshua Bengio · Sungjin Ahn

Due to the inherent model uncertainty, learning to infer Bayesian posterior from a few-shot dataset is an important step towards robust meta-learning. In this paper, we propose a novel Bayesian model-agnostic meta-learning method. The proposed method combines efficient gradient-based meta-learning with nonparametric variational inference in a principled probabilistic framework. Unlike previous methods, during fast adaptation, the method is capable of learning complex uncertainty structure beyond a simple Gaussian approximation, and during meta-update, a novel Bayesian mechanism prevents meta-level overfitting. Remaining a gradient-based method, it is also the first Bayesian model-agnostic meta-learning method applicable to various tasks including reinforcement learning. Experiment results show the accuracy and robustness of the proposed method in sinusoidal regression, image classification, active learning, and reinforcement learning.

Differentially Private k-Means with Constant Multiplicative Error

Spotlight | Tue Dec 4th 03:40 -- 03:45 PM @ Room 517 CD

Uri Stemmer · Haim Kaplan

We design new differentially private algorithms for the Euclidean k-means problem, both in the centralized model and in the local model of differential privacy. In both models, our algorithms achieve significantly improved error guarantees than the previous state-of-the-art. In addition, in the local model, our algorithm significantly reduces the number of interaction rounds. Although the problem has been widely studied in the context of differential privacy, all of the existing constructions achieve only super constant approximation factors. We present, for the first time, efficient private algorithms for the problem with constant multiplicative error. Furthermore, we show how to modify our algorithms so they compute private coresets for k-means clustering in both models.

Learning to Optimize Tensor Programs

Spotlight | Tue Dec 4th 03:45 -- 03:50 PM @ Room 220 CD

Tianqi Chen · Lianmin Zheng · Eddie Yan · Ziheng Jiang · Thierry Moreau · Luis Ceze · Carlos Guestrin · Arvind Krishnamurthy

We introduce a learning-based framework to optimize tensor programs for deep learning workloads. Efficient implementations of tensor operators, such as matrix multiplication and high dimensional convolution are key enablers of effective deep learning systems. However, existing systems rely on manually optimized libraries such as cuDNN where only a narrow range of server class GPUs are well-supported. The reliance on hardware specific operator libraries limits the applicability of high-level graph optimizations and incurs significant engineering costs when deploying to new hardware targets. We use learning to remove this engineering burden. We learn domain specific statistical cost models to guide the search of tensor operator implementations over billions of possible program variants. We further accelerate the search by effective model transfer across workloads. Experimental results show that our framework delivers performance competitive with state-of-the-art hand-tuned libraries for low-power CPU, mobile GPU, and server-class GPU.

Probabilistic Neural Programmed Networks for Scene Generation

Spotlight | Tue Dec 4th 03:45 -- 03:50 PM @ Room 220 E

Zhiwei Deng · Jiacheng Chen · YIFANG FU · Greg Mori

In this paper we address the text to scene image generation problem. Generative models that capture the variability in complicated scenes containing rich semantics is a grand goal of image generation. Complicated scene images contain rich visual elements, compositional visual concepts, and complicated relations between objects. Generative models, as an analysis-by-synthesis process, should encompass the following three core components: 1) the generation process that composes the scene; 2) what are the primitive visual elements and how are they composed; 3) the rendering of abstract concepts into their pixel-level realizations. We propose PNP-Net, a variational auto-encoder framework that addresses these three challenges: it flexibly composes images with a dynamic network structure, learns a set of distribution transformers that can compose distributions based on semantics, and decodes samples from these distributions into realistic images.

A Spectral View of Adversarially Robust Features

Spotlight | Tue Dec 4th 03:45 -- 03:50 PM @ Room 517 CD

Shivam Garg · Vatsal Sharan · Brian Zhang · Gregory Valiant

Given the apparent difficulty of learning models that are robust to adversarial perturbations, we propose tackling the simpler problem of developing adversarially robust features. Specifically, given a dataset and metric of interest, the goal is to return a function (or multiple functions) that 1) is robust to adversarial perturbations, and 2) has significant variation across the datapoints. We establish strong connections between adversarially robust features and a natural spectral property of the geometry of the dataset and metric of interest. This connection can be leveraged to provide both robust features, and a lower bound on the robustness of any function that has significant variance across the dataset. Finally, we provide empirical evidence that the adversarially robust features given by this spectral approach can be fruitfully leveraged to learn a robust (and accurate) model.

Generalisation of structural knowledge in the hippocampal-

entorhinal system

Oral | Tue Dec 4th 03:50 -- 04:05 PM @ Room 220 CD

James Whittington · Timothy Muller · Shirely Mark · Caswell Barry · Tim Behrens

A central problem to understanding intelligence is the concept of generalisation. This allows previously learnt structure to be exploited to solve tasks in novel situations differing in their particularities. We take inspiration from neuroscience, specifically the hippocampal-entorhinal system known to be important for generalisation. We propose that to generalise structural knowledge, the representations of the structure of the world, i.e. how entities in the world relate to each other, need to be separated from representations of the entities themselves. We show, under these principles, artificial neural networks embedded with hierarchy and fast Hebbian memory, can learn the statistics of memories and generalise structural knowledge. Spatial neuronal representations mirroring those found in the brain emerge, suggesting spatial cognition is an instance of more general organising principles. We further unify many entorhinal cell types as basis functions for constructing transition graphs, and show these representations effectively utilise memories. We experimentally support model assumptions, showing a preserved relationship between entorhinal grid and hippocampal place cells across environments.

Neural Ordinary Differential Equations

Oral | Tue Dec 4th 03:50 -- 04:05 PM @ Room 220 E

Tian Qi Chen · Yulia Rubanova · Jesse Bettencourt · David Duvenaud

We introduce a new family of deep neural network models. Instead of specifying a discrete sequence of hidden layers, we parameterize the derivative of the hidden state using a neural network. The output of the network is computed using a blackbox differential equation solver. These continuous-depth models have constant memory cost, adapt their evaluation strategy to each input, and can explicitly trade numerical precision for speed. We demonstrate these properties in continuous-depth residual networks and continuous-time latent variable models. We also construct continuous normalizing flows, a generative model that can train by maximum likelihood, without partitioning or ordering the data dimensions. For training, we show how to scalably backpropagate through any ODE solver, without access to its internal operations. This allows end-to-end training of ODEs within larger models.

Model-Agnostic Private Learning

Oral | Tue Dec 4th 03:50 -- 04:05 PM @ Room 517 CD

Raef Bassily · Abhradeep Guha Thakurta · Om Dipakbhai Thakkar

We design differentially private learning algorithms that are agnostic to the learning model assuming access to limited amount of unlabeled public data. First, we give a new differentially private algorithm for answering a sequence of m online classification queries (given by a sequence of m unlabeled public feature vectors) based on a private training set. Our private algorithm follows the paradigm of subsample-and-aggregate, in which any generic non-private learner is trained on disjoint subsets of the private training set, then for each classification query, the votes of the resulting classifiers ensemble are aggregated in a differentially private fashion. Our private aggregation is based on a novel combination of distance-to-instability framework [Smith & Thakurta 2013] and the sparse-vector technique [Dwork et al. 2009, Hardt & Talwar 2010]. We show that our algorithm makes a conservative use of the privacy budget. In particular, if the underlying non-private learner yields classification error at most $\alpha\in (0, 1)$, then our construction answers more queries, by at least a factor of $1/\alpha$ in some cases, than what is implied by a straightforward application of the advanced composition theorem for differential privacy. Next, we apply the knowledge transfer technique to construct a private learner that outputs a classifier, which can be used to answer unlimited number of queries. In the PAC model, we analyze our construction and prove upper bounds on the sample complexity for both the realizable and the non-realizable cases. As in non-private sample complexity, our bounds are completely characterized by the VC dimension of the concept class.

A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks

Spotlight | Tue Dec 4th 04:05 -- 04:10 PM @ Room 220 CD

Jeffrey Chan · Valerio Perrone · Jeffrey Spence · Paul Jenkins · Sara Mathieson · Yun Song

An explosion of high-throughput DNA sequencing in the past decade has led to a surge of interest in population-scale inference with whole-genome data. Recent work in population genetics has centered on designing inference methods for relatively simple model classes, and few scalable general-purpose inference techniques exist for more realistic, complex models. To achieve this, two inferential challenges need to be addressed: (1) population data are exchangeable, calling for methods that efficiently exploit the symmetries of the data, and (2) computing likelihoods is intractable as it requires integrating over a set of correlated, extremely high-dimensional latent variables. These challenges are traditionally tackled by likelihood-free methods that use scientific simulators to generate datasets and reduce them to hand-designed, permutation-invariant summary statistics, often leading to inaccurate inference. In this work, we develop an exchangeable neural network that performs summary statistic-free, likelihood-free inference. Our framework can be applied in a black-box fashion across a variety of simulation-based tasks, both within and outside biology. We demonstrate the power of our approach on the recombination hotspot testing problem, outperforming the state-of-the-art.

Bias and Generalization in Deep Generative Models: An Empirical Study

Spotlight | Tue Dec 4th 04:05 -- 04:10 PM @ Room 220 E

Shengjia Zhao · Hongyu Ren · Arianna Yuan · Jiaming Song · Noah Goodman · Stefano Ermon

In high dimensional settings, density estimation algorithms rely crucially on their inductive bias. Despite recent empirical success, the inductive bias of deep generative models is not well understood. In this paper we propose a framework to systematically investigate bias and generalization in deep generative models of images by probing the learning algorithm with carefully designed training datasets. By measuring properties of the learned distribution, we are able to find interesting patterns of generalization. We verify that these patterns are consistent across datasets, common models and architectures.

Bounded-Loss Private Prediction Markets

Spotlight | Tue Dec 4th 04:05 -- 04:10 PM @ Room 517 CD

Rafael Frongillo · Bo Waggoner

Prior work has investigated variations of prediction markets that preserve participants' (differential) privacy, which formed the basis of useful mechanisms for purchasing data for machine learning objectives. Such markets required potentially unlimited financial subsidy, however, making them impractical. In this work, we design an adaptively-growing prediction market with a bounded financial subsidy, while achieving privacy, incentives to produce accurate predictions, and precision in the sense that market prices are not heavily impacted by the added privacy-preserving noise. We briefly discuss how our mechanism can extend to the data-purchasing setting, and its relationship to traditional learning algorithms.

Generalizing Tree Probability Estimation via Bayesian Networks

Spotlight | Tue Dec 4th 04:10 -- 04:15 PM @ Room 220 CD

Cheng Zhang · Frederick A Matsen IV

Probability estimation is one of the fundamental tasks in statistics and machine learning. However, standard methods for probability estimation on discrete objects do not handle object structure in a satisfactory manner. In this paper, we derive a general Bayesian network formulation for probability estimation on leaf-labeled trees that enables flexible approximations which can generalize beyond observations. We show that efficient algorithms for learning Bayesian networks can be easily extended to probability estimation on this challenging structured space. Experiments on both synthetic and real data show that our methods greatly outperform the current practice of using the empirical distribution, as well as a previous effort for probability estimation on trees.

Robustness of conditional GANs to noisy labels

Spotlight | Tue Dec 4th 04:10 -- 04:15 PM @ Room 220 E

Kiran Thekumparampil · Ashish Khetan · Zinan Lin · Sewoong Oh

We study the problem of learning conditional generators from noisy labeled samples, where the labels are corrupted by random noise. A standard training of conditional GANs will not only produce samples with wrong labels, but also generate poor quality samples. We consider two scenarios, depending on whether the noise model is known or not. When the distribution of the noise is known, we introduce a novel architecture which we call Robust Conditional GAN (RCGAN). The main idea is to corrupt the label of the generated sample before feeding to the adversarial discriminator, forcing the generator to produce samples with clean labels. This approach of passing through a matching noisy channel is justified by accompanying multiplicative approximation bounds between the loss of the RCGAN and the distance between the clean real distribution and the generator distribution. This shows that the proposed approach is robust, when used with a carefully chosen discriminator architecture, known as projection discriminator. When the distribution of the noise is not known, we provide an extension of our architecture, which we call RCGAN-U, that learns the noise model simultaneously while training the generator. We show experimentally on MNIST and CIFAR-10 datasets that both the approaches consistently improve upon baseline approaches, and RCGAN-U closely matches the performance of RCGAN.

---

cpSGD: Communication-efficient and differentially-private distributed SGD

Spotlight | Tue Dec 4th 04:10 -- 04:15 PM @ Room 517 CD

Naman Agarwal · Ananda Theertha Suresh · Felix Xinnan Yu · Sanjiv Kumar · Brendan McMahan

Distributed stochastic gradient descent is an important subroutine in distributed learning. A setting of particular interest is when the clients are mobile devices, where two important concerns are communication efficiency and the privacy of the clients. Several recent works have focused on reducing the communication cost or introducing privacy guarantees, but none of the proposed communication efficient methods are known to be privacy preserving and none of the known privacy mechanisms are known to be communication efficient. To this end, we study algorithms that achieve both communication efficiency and differential privacy. For d variables and $n \approx d$ clients, the proposed method uses $\cO(\log \log(nd))$ bits of communication per client per coordinate and ensures constant privacy.

We also improve previous analysis of the \emph{Binomial mechanism} showing that it achieves nearly the same utility as the Gaussian mechanism, while requiring fewer representation bits, which can be of independent interest.

Geometry Based Data Generation

Spotlight | Tue Dec 4th 04:15 -- 04:20 PM @ Room 220 CD

Ofir Lindenbaum · Jay Stanley · Guy Wolf · Smita Krishnaswamy

We propose a new type of generative model for high-dimensional data that learns a manifold geometry of the data, rather than density, and can generate points evenly along this manifold. This is in contrast to existing generative models that represent data density, and are strongly affected by noise and other artifacts of data collection. We demonstrate how this approach corrects sampling biases and artifacts, thus improves several downstream data analysis tasks, such as clustering and classification. Finally, we demonstrate that this approach is especially useful in biology where, despite the advent of single-cell technologies, rare subpopulations and gene-interaction relationships are affected by biased sampling. We show that SUGAR can generate hypothetical populations, and it is able to reveal intrinsic patterns and mutual-information relationships between genes on a single-cell RNA sequencing dataset of hematopoiesis.

BourGAN: Generative Networks with Metric Embeddings

Spotlight | Tue Dec 4th 04:15 -- 04:20 PM @ Room 220 E

Chang Xiao · Peilin Zhong · Changxi Zheng

This paper addresses the mode collapse for generative adversarial networks (GANs). We view modes as a geometric structure of data distribution in a metric space. Under this geometric lens, we embed subsamples of the dataset from an arbitrary metric space into the L2 space, while preserving their pairwise distance distribution. Not only does this metric embedding determine the dimensionality of the latent space automatically, it also enables us to construct a mixture of Gaussians to draw latent space random vectors. We use the Gaussian mixture model in tandem with a simple augmentation of the objective function to train GANs. Every major step of our method is supported by theoretical analysis, and our experiments on real and synthetic data confirm that the generator is able to produce samples spreading over most of the modes while avoiding unwanted samples, outperforming several recent GAN variants on a number of metrics and offering new features.

Adversarially Robust Generalization Requires More Data

Spotlight | Tue Dec 4th 04:15 -- 04:20 PM @ Room 517 CD

Ludwig Schmidt · Shibani Santurkar · Dimitris Tsipras · Kunal Talwar · Aleksander Madry

Machine learning models are often susceptible to adversarial perturbations of their inputs. Even small perturbations can cause state-of-the-art classifiers with high "standard" accuracy to produce an incorrect prediction with high confidence. To better understand this phenomenon, we study adversarially robust learning from the viewpoint of generalization. We show that already in a simple natural data model, the sample complexity of robust learning can be significantly larger than that of "standard" learning. This gap is information theoretic and holds irrespective of the training algorithm or the model family. We complement our theoretical results with experiments on popular image classification datasets and show that a similar gap exists here as well. We postulate that the difficulty of training robust classifiers stems, at least partially, from this inherently larger sample complexity.

Point process latent variable models of larval zebrafish behavior

Spotlight | Tue Dec 4th 04:20 -- 04:25 PM @ Room 220 CD

Anuj Sharma · Scott Linderman · Robert Johnson · Florian Engert

A fundamental goal of systems neuroscience is to understand how neural activity gives rise to natural behavior. In order to achieve this goal, we must first build comprehensive models that offer quantitative descriptions of behavior. We develop a new class of probabilistic models to tackle this challenge in the study of larval zebrafish, an important model organism for neuroscience. Larval zebrafish locomote via sequences of punctate swim bouts--brief flicks of the tail--which are naturally modeled as a marked point process. However, these sequences of swim bouts belie a set of discrete and continuous internal states, latent variables that are not captured by standard point process models. We incorporate these variables as latent marks of a point process and explore various models for their dynamics. To infer the latent variables and fit the parameters of this model, we develop an amortized variational inference algorithm that targets the collapsed posterior distribution, analytically marginalizing out the discrete latent variables. With a dataset of over 120,000 swim bouts, we show that our models reveal interpretable discrete classes of swim bouts and continuous internal states like hunger that modulate their dynamics. These models are a major step toward understanding the natural behavioral program of the larval zebrafish and, ultimately, its neural underpinnings.

Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Spotlight | Tue Dec 4th 04:20 -- 04:25 PM @ Room 220 E

Timur Garipov · Pavel Izmailov · Dmitrii Podoprikhin · Dmitry Vetrov · Andrew Wilson

The loss functions of deep neural networks are complex and their geometric properties are not well understood. We show that the optima of these complex loss functions are in fact connected by simple curves, over which training and test accuracy are nearly constant. We introduce a training procedure to discover these high-accuracy pathways between modes. Inspired by this new geometric insight, we also propose a new ensembling method entitled Fast Geometric Ensembling (FGE). Using FGE we can train high-performing ensembles in the time required to train a single model. We achieve improved performance compared to the recent state-of-the-art Snapshot Ensembles, on CIFAR-10, CIFAR-100, and ImageNet.

Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples

Spotlight | Tue Dec 4th 04:20 -- 04:25 PM @ Room 517 CD

Guanhong Tao · Shiqing Ma · Yingqi Liu · Xiangyu Zhang

Adversarial sample attacks perturb benign inputs to induce DNN misbehaviors. Recent research has demonstrated the widespread presence and the devastating consequences of such attacks. Existing defense techniques either assume prior knowledge of specific attacks or may not work well on complex models due to their underlying assumptions. We argue that adversarial sample attacks are deeply entangled with interpretability of DNN models: while classification results on benign inputs can be reasoned based on the human perceptible features/attributes, results on adversarial samples can hardly be explained. Therefore, we propose a novel adversarial sample detection technique for face recognition models, based on interpretability. It features a novel bi-directional correspondence inference between attributes and internal neurons to identify neurons critical for individual attributes. The activation values of critical neurons are enhanced to amplify the reasoning part of the computation and the values of other neurons are weakened to suppress the uninterpretable part. The classification results after such transformation are compared with those of the original model to detect adversaries. Results show that our technique can achieve 94% detection accuracy for 7 different kinds of attacks with 9.91% false positives on benign inputs. In contrast, a state-of-the-art feature squeezing technique can only achieve 55% accuracy with 23.3% false positives.

A probabilistic population code based on neural samples

Oral | Tue Dec 4th 04:25 -- 04:40 PM @ Room 220 CD

Sabyasachi Shivkumar · Richard Lange · Ankani Chattoraj · Ralf Haefner

Sensory processing is often characterized as implementing probabilistic inference: networks of neurons compute posterior beliefs over unobserved causes given the sensory inputs. How these beliefs are computed and represented by neural responses is much-debated (Fiser et al. 2010, Pouget et al. 2013). A central debate concerns the question of whether neural responses represent samples of latent variables (Hoyer & Hyvarinnen 2003) or parameters of their distributions (Ma et al. 2006) with efforts being made to distinguish between them (Grabska-Barwinska et al. 2013). A separate debate addresses the question of whether neural responses are proportionally related to the encoded probabilities (Barlow 1969), or proportional to the logarithm of those probabilities (Jazayeri & Movshon 2006, Ma et al. 2006, Beck et al. 2012). Here, we show that these alternatives -- contrary to common assumptions -- are not mutually exclusive and that the very same system can be compatible with all of them. As a central analytical result, we show that modeling neural responses in area V1 as samples from a posterior distribution over latents in a linear Gaussian model of the image implies that those neural responses form a linear Probabilistic Population Code (PPC, Ma et al. 2006). In particular, the posterior distribution over some experimenter-defined variable like "orientation" is part of the exponential family with sufficient statistics that are linear in the neural sampling-based firing rates.

How Does Batch Normalization Help Optimization?

Oral | Tue Dec 4th 04:25 -- 04:40 PM @ Room 220 E

Shibani Santurkar · Dimitris Tsipras · Andrew Ilyas · Aleksander Madry

Batch Normalization (BatchNorm) is a widely adopted technique that enables faster and more stable training of deep neural networks (DNNs). Despite its pervasiveness, the exact reasons for BatchNorm's effectiveness are still poorly understood. The popular belief is that this effectiveness stems from controlling the change of the layers' input distributions during training to reduce the so-called "internal covariate shift". In this work, we demonstrate that such distributional stability of layer inputs has little to do with the success of BatchNorm. Instead, we uncover a more fundamental impact of BatchNorm on the training process: it makes the optimization landscape significantly smoother. This smoothness induces a more predictive and stable behavior of the gradients, allowing for faster training.

Learning to Solve SMT Formulas

Oral | Tue Dec 4th 04:25 -- 04:40 PM @ Room 517 CD

Mislav Balunovic · Pavol Bielik · Martin Vechev

We present a new approach for learning to solve SMT formulas. We phrase the challenge of solving SMT formulas as a tree search problem where at each step a transformation is applied to the input formula until the formula is solved. Our approach works in two phases: first, given a dataset of unsolved formulas we learn a policy that for each formula selects a suitable transformation to apply at each step in order to solve the formula, and second, we synthesize a strategy in the form of a loop-free program with branches. This strategy is an interpretable representation of the policy decisions and is used to guide the SMT solver to decide formulas more efficiently, without requiring any modification to the solver itself and without needing to evaluate the learned policy at inference time. We show that our approach is effective in practice - it solves 17% more formulas over a range of benchmarks and achieves up to 100x runtime improvement over a state-of-the-art SMT solver.

Sparse Attentive Backtracking: Temporal Credit Assignment Through Reminding

Spotlight | Tue Dec 4th 04:40 -- 04:45 PM @ Room 220 CD

Nan Rosemary Ke · Anirudh Goyal ALIAS PARTH GOYAL · Olexa Bilaniuk · Jonathan Binas · Michael Mozer · Chris Pal · Yoshua Bengio

Learning long-term dependencies in extended temporal sequences requires credit assignment to events far back in the past. The most common method for training recurrent neural networks, back-propagation through time (BPTT), requires credit information to be propagated backwards through every single step of the forward computation, potentially over thousands or millions of time steps. This becomes computationally expensive or even infeasible when used with long sequences. Importantly, biological brains are unlikely to perform such detailed reverse replay over very long sequences of internal states (consider days, months, or years.) However, humans are often reminded of past memories or mental states which are associated with the current mental state. We consider the hypothesis that such memory associations between past and present could be used for credit assignment through arbitrarily long sequences, propagating the credit assigned to the current state to the associated past state. Based on this principle, we study a novel algorithm which only back-propagates through a few of these temporal skip connections, realized by a learned attention mechanism that associates current states with relevant past states. We demonstrate in experiments that our method matches or outperforms regular BPTT and truncated BPTT in tasks involving particularly long-term dependencies, but without requiring the biologically implausible backward replay through the whole history of states. Additionally, we demonstrate that the proposed method transfers to longer sequences significantly better than LSTMs trained with BPTT and LSTMs trained with full self-attention.

Training Neural Networks Using Features Replay

Spotlight | Tue Dec 4th 04:40 -- 04:45 PM @ Room 220 E

Zhouyuan Huo · Bin Gu · Heng Huang

Training a neural network using backpropagation algorithm requires passing error gradients sequentially through the network. The backward locking prevents us from updating network layers in parallel and fully leveraging the computing resources. Recently, there are several works trying to decouple and parallelize the backpropagation algorithm. However, all of them suffer from severe accuracy loss or memory explosion when the neural network is deep. To address these challenging issues, we propose a novel parallel-objective formulation for the objective function of the neural network. After that, we introduce features replay algorithm and prove that it is guaranteed to converge to critical points for the non-convex problem under certain conditions. Finally, we apply our method to training deep convolutional neural networks, and the experimental results show that the proposed method achieves {faster} convergence, {lower} memory consumption, and {better} generalization error than compared methods.

Towards Robust Detection of Adversarial Examples

Spotlight | Tue Dec 4th 04:40 -- 04:45 PM @ Room 517 CD

Tianyu Pang · Chao Du · Yinpeng Dong · Jun Zhu

Although the recent progress is substantial, deep learning methods can be vulnerable to the maliciously generated adversarial examples. In this paper, we present a novel training procedure and a thresholding test strategy, towards robust detection of adversarial examples. In training, we propose to minimize the reverse cross-entropy (RCE), which encourages a deep network to learn latent representations that better distinguish adversarial examples from normal ones. In testing, we propose to use a thresholding strategy as the detector to filter out adversarial examples for reliable predictions. Our method is simple to implement using standard algorithms, with little extra training cost compared to the common cross-entropy minimization. We apply our method to defend various attacking methods on the widely used MNIST and CIFAR-10 datasets, and achieve significant improvements on robust predictions under all the threat models in the adversarial setting.

Learning Temporal Point Processes via Reinforcement Learning

Spotlight | Tue Dec 4th 04:45 -- 04:50 PM @ Room 220 CD

Shuang Li · Shuai Xiao · Shixiang Zhu · Nan Du · Yao Xie · Le Song

Social goods, such as healthcare, smart city, and information networks, often produce ordered event data in continuous time. The generative processes of these event data can be very complex, requiring flexible models to capture their dynamics. Temporal point processes offer an elegant framework for modeling event data without discretizing the time. However, the existing maximum-likelihood-estimation (MLE) learning paradigm requires hand-crafting the intensity function beforehand and cannot directly monitor the goodness-of-fit of the estimated model in the process of training. To alleviate the risk of model-misspecification in MLE, we propose to generate samples from the generative model and monitor the quality of the samples in the process of training until the samples and the real data are indistinguishable. We take inspiration from reinforcement learning (RL) and treat the generation of each event as the action taken by a stochastic policy. We parameterize the policy as a flexible recurrent neural network and gradually improve the policy to mimic the observed event distribution. Since the reward function is unknown in this setting, we uncover an analytic and nonparametric form of the reward function using an inverse reinforcement learning formulation. This new RL framework allows us to derive an efficient policy gradient algorithm for learning flexible point process models, and we show that it performs well in both synthetic and real data.

Step Size Matters in Deep Learning

Spotlight | Tue Dec 4th 04:45 -- 04:50 PM @ Room 220 E

Kamil Nar · Shankar Sastry

Training a neural network with the gradient descent algorithm gives rise to a discrete-time nonlinear dynamical system. Consequently, behaviors that are typically observed in these systems emerge during training, such as convergence to an orbit but not to a fixed point or dependence of convergence on the initialization. Step size of the algorithm plays a critical role in these behaviors: it determines the subset of the local optima that the algorithm can converge to, and it specifies the magnitude of the oscillations if the algorithm converges to an orbit. To elucidate the effects of the step size on training of neural networks, we study the gradient descent algorithm as a discrete-time dynamical system, and by analyzing the Lyapunov stability of different solutions, we show the relationship between the step size of the algorithm and the solutions that can be obtained with this algorithm. The results provide an explanation for several phenomena observed in practice, including the deterioration in the training error with increased depth, the hardness of estimating linear mappings with large singular values, and the distinct performance of deep residual networks.

Neural Architecture Search with Bayesian Optimisation and Optimal Transport

Spotlight | Tue Dec 4th 04:45 -- 04:50 PM @ Room 517 CD

Kirthevasan Kandasamy · Willie Neiswanger · Jeff Schneider · Barnabas Poczos · Eric Xing

Bayesian Optimisation (BO) refers to a class of methods for global optimisation of a function f which is only accessible via point evaluations. It is typically used in settings where f is expensive to evaluate. A common use case for BO in machine learning is model selection, where it is not possible to analytically model the generalisation performance of a statistical model, and we resort to noisy and expensive training and validation procedures to choose the best model. Conventional BO methods have focused on Euclidean and categorical domains, which, in the context of model selection, only permits tuning scalar hyper-parameters of machine learning algorithms. However, with the surge of interest in deep learning, there is an increasing demand to tune neural network architectures. In this work, we develop NASBOT, a Gaussian process based BO framework for neural architecture search. To accomplish this, we develop a distance metric in the space of neural network architectures which can be computed efficiently via an optimal transport program. This distance might be of independent interest to the deep learning community as it may find applications outside of BO. We demonstrate that NASBOT outperforms other alternatives for architecture search in several cross validation based model selection tasks on multi-layer perceptrons and convolutional neural networks.

Precision and Recall for Time Series

Spotlight | Tue Dec 4th 04:50 -- 04:55 PM @ Room 220 CD

Nesime Tatbul · Tae Jun Lee · Stan Zdonik · Mejbah Alam · Justin Gottschlich

Classical anomaly detection is principally concerned with point-based anomalies, those anomalies that occur at a single point in time. Yet, many real-world anomalies are range-based, meaning they occur over a period of time. Motivated by this observation, we present a new mathematical model to evaluate the accuracy of time series classification algorithms. Our model expands the well-known Precision and Recall metrics to measure ranges, while simultaneously enabling customization support for domain-specific preferences.

Neural Tangent Kernel: Convergence and Generalization in

Neural Networks

Spotlight | Tue Dec 4th 04:50 -- 04:55 PM @ Room 220 E

Arthur Jacot-Guillarmod · Clement Hongler · Franck Gabriel

At initialization, artificial neural networks (ANNs) are equivalent to Gaussian processes in the infinite-width limit, thus connecting them to kernel methods. We prove that the evolution of an ANN during training can also be described by a kernel: during gradient descent on the parameters of an ANN, the network function (which maps input vectors to output vectors) follows the so-called kernel gradient associated with a new object, which we call the Neural Tangent Kernel (NTK). This kernel is central to describe the generalization features of ANNs. While the NTK is random at initialization and varies during training, in the infinite-width limit it converges to an explicit limiting kernel and stays constant during training. This makes it possible to study the training of ANNs in function space instead of parameter space. Convergence of the training can then be related to the positive-definiteness of the limiting NTK. We then focus on the setting of least-squares regression and show that in the infinite-width limit, the network function follows a linear differential equation during training. The convergence is fastest along the largest kernel principal components of the input data with respect to the NTK, hence suggesting a theoretical motivation for early stopping. Finally we study the NTK numerically, observe its behavior for wide networks, and compare it to the infinite-width limit.

Data-Driven Clustering via Parameterized Lloyd's Families

Spotlight | Tue Dec 4th 04:50 -- 04:55 PM @ Room 517 CD

Maria-Florina Balcan · Travis Dick · Colin White

Algorithms for clustering points in metric spaces is a long-studied area of research. Clustering has seen a multitude of work both theoretically, in understanding the approximation guarantees possible for many objective functions such as k-median and k-means clustering, and experimentally, in finding the fastest algorithms and seeding procedures for Lloyd's algorithm. The performance of a given clustering algorithm depends on the specific application at hand, and this may not be known up front. For example, a "typical instance" may vary depending on the application, and different clustering heuristics perform differently depending on the instance. In this paper, we define an infinite family of algorithms generalizing Lloyd's algorithm, with one parameter controlling the the initialization procedure, and another parameter controlling the local search procedure. This family of algorithms includes the celebrated k-means++ algorithm, as well as the classic farthest-first traversal algorithm. We design efficient learning algorithms which receive samples from an application-specific distribution over clustering instances and learn a near-optimal clustering algorithm from the class. We show the best parameters vary significantly across datasets such as MNIST, CIFAR, and mixtures of Gaussians. Our learned algorithms never perform worse than k-means++, and on some datasets we see significant improvements.

Bayesian Nonparametric Spectral Estimation

Spotlight | Tue Dec 4th 04:55 -- 05:00 PM @ Room 220 CD

Felipe Tobar

Spectral estimation (SE) aims to identify how the energy of a signal (e.g., a time series) is distributed across different frequencies. This can become particularly challenging when only partial and noisy observations of the signal are available, where current methods fail to handle uncertainty appropriately. In this context, we propose a joint probabilistic model for signals, observations and spectra, where SE is addressed as an inference problem. Assuming a Gaussian process prior over the signal, we apply Bayes' rule to find the analytic posterior distribution of the spectrum given a set of observations. Besides its expressiveness and natural account of spectral uncertainty, the proposed model also provides a functional-form representation of the power spectral density, which can be optimised efficiently. Comparison with previous approaches is addressed theoretically, showing that the proposed method is an infinite-dimensional variant of the Lomb-Scargle approach, and also empirically through three experiments.

Hierarchical Graph Representation Learning with Differentiable Pooling

Spotlight | Tue Dec 4th 04:55 -- 05:00 PM @ Room 220 E

Zhitao Ying · Jiaxuan You · Christopher Morris · Xiang Ren · Will Hamilton · Jure Leskovec

Recently, graph neural networks (GNNs) have revolutionized the field of graph representation learning through effectively learned node embeddings, and achieved state-of-the-art results in tasks such as node classification and link prediction. However, current GNN methods are inherently flat and do not learn hierarchical representations of graphs---a limitation that is especially problematic for the task of graph classification, where the goal is to predict the label associated with an entire graph. Here we propose DiffPool, a differentiable graph pooling module that can generate hierarchical representations of graphs and can be combined with various graph neural network architectures in an end-to-end fashion. DiffPool learns a differentiable soft cluster assignment for nodes at each layer of a deep GNN, mapping nodes to a set of clusters, which then form the coarsened input for the next GNN layer. Our experimental results show that combining existing GNN methods with DiffPool yields an average improvement of 5-10% accuracy on graph classification benchmarks, compared to all existing pooling approaches, achieving a new state-of-the-art on four out of five benchmark datasets.

Supervising Unsupervised Learning

Spotlight | Tue Dec 4th 04:55 -- 05:00 PM @ Room 517 CD

Vikas Garg · Adam Kalai

We introduce a framework to transfer knowledge acquired from a repository of (heterogeneous) supervised datasets to new unsupervised datasets. Our perspective avoids the subjectivity inherent in unsupervised learning by reducing it to supervised learning, and provides a principled way to evaluate unsupervised algorithms. We demonstrate the versatility of our framework via rigorous agnostic bounds on a variety of unsupervised problems. In the context of clustering, our approach helps choose the number of clusters and the clustering algorithm, remove the outliers, and provably circumvent Kleinberg's impossibility result. Experiments across hundreds of problems demonstrate improvements in performance on unsupervised data with simple algorithms despite the fact our problems come from heterogeneous domains. Additionally, our framework lets us leverage deep networks to learn common features across many small datasets, and perform zero shot learning.

Point process latent variable models of larval zebrafish behavior

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #1

Anuj Sharma · Scott Linderman · Robert Johnson · Florian Engert

A fundamental goal of systems neuroscience is to understand how neural activity gives rise to natural behavior. In order to achieve this goal, we must first build comprehensive models that offer quantitative descriptions of behavior. We develop a new class of probabilistic models to tackle this challenge in the study of larval zebrafish, an important model organism for neuroscience. Larval zebrafish locomote via sequences of punctate swim bouts--brief flicks of the tail--which are naturally modeled as a marked point process. However, these sequences of swim bouts belie a set of discrete and continuous internal states, latent variables that are not captured by standard point process models. We incorporate these variables as latent marks of a point process and explore various models for their dynamics. To infer the latent variables and fit the parameters of this model, we develop an amortized variational inference algorithm that targets the collapsed posterior distribution, analytically marginalizing out the discrete latent variables. With a dataset of over 120,000 swim bouts, we show that our models reveal interpretable discrete classes of swim bouts and continuous internal states like hunger that modulate their dynamics. These models are a major step toward understanding the natural behavioral program of the larval zebrafish and, ultimately, its neural underpinnings.

Provably Correct Automatic Sub-Differentiation for Qualified Programs

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #2

Sham Kakade · Jason Lee

The \emph{Cheap Gradient Principle}~\citep{Griewank:2008:EDP:1455489} --- the computational cost of computing a \$d\$-dimensional vector of partial derivatives of a scalar function is nearly the same (often within a factor of \$5\$) as that of simply computing the scalar function itself --- is of central importance in optimization; it allows us to quickly obtain (high-dimensional) gradients of scalar loss functions which are subsequently used in black box gradient-based optimization procedures. The current state of affairs is markedly different with regards to computing sub-derivatives: widely used ML libraries, including TensorFlow and PyTorch, do \emph{not} correctly compute (generalized) sub-derivatives even on simple differentiable examples. This work considers the question: is there a \emph{Cheap Sub-gradient Principle}? Our main result shows that, under certain restrictions on our library of non-smooth functions (standard in non-linear programming), provably correct generalized sub-derivatives can be computed at a computational cost that is within a (dimension-free) factor of \$6\$ of the cost of computing the scalar function itself.

Neural Ordinary Differential Equations

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #3

Tian Qi Chen · Yulia Rubanova · Jesse Bettencourt · David Duvenaud

We introduce a new family of deep neural network models. Instead of specifying a discrete sequence of hidden layers, we parameterize the derivative of the hidden state using a neural network. The output of the network is computed using a blackbox differential equation solver. These continuous-depth models have constant memory cost, adapt their evaluation strategy to each input, and can explicitly trade numerical precision for speed. We demonstrate these properties in continuous-depth residual networks and continuous-time latent variable models. We also construct continuous normalizing flows, a generative model that can train by maximum likelihood, without partitioning or ordering the data dimensions. For training, we show how to scalably backpropagate through any ODE solver, without access to its internal operations. This allows end-to-end training of ODEs within larger models.

Information Constraints on Auto-Encoding Variational Bayes

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #4

Romain Lopez · Jeffrey Regier · Michael Jordan · Nir Yosef

Parameterizing the approximate posterior of a generative model with neural networks has become a common theme in recent machine learning research. While providing appealing flexibility, this approach makes it difficult to impose or assess structural constraints such as conditional independence. We propose a framework for learning representations that relies on Auto-Encoding Variational Bayes and whose search space is constrained via kernel-based measures of independence. In particular, our method employs the d-variable Hilbert-Schmidt Independence Criterion (dHSIC) to enforce independence between the latent representations and arbitrary nuisance factors. We show how to apply this method to a range of problems, including the problems of learning invariant representations and the learning of interpretable representations. We also present a full-fledged application to single-cell RNA sequencing (scRNA-seq). In this setting the biological signal in mixed in complex ways with sequencing errors and sampling effects. We show that our method out-performs the state-of-the-art in this domain.

Robustness of conditional GANs to noisy labels

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #5

Kiran Thekumparampil · Ashish Khetan · Zinan Lin · Sewoong Oh

We study the problem of learning conditional generators from noisy labeled samples, where the labels are corrupted by random noise. A standard training of conditional GANs will not only produce samples with wrong labels, but also generate poor quality samples. We consider two scenarios, depending on whether the noise model is known or not. When the distribution of the noise is known, we introduce a novel architecture which we call Robust Conditional GAN (RCGAN). The main idea is to corrupt the label of the generated sample before feeding to the adversarial discriminator, forcing the generator to produce samples with clean labels. This approach of passing through a matching noisy channel is justified by accompanying multiplicative approximation bounds between the loss of the RCGAN and the distance between the clean real distribution and the generator distribution. This shows that the proposed approach is robust, when used with a carefully chosen discriminator architecture, known as projection discriminator. When the distribution of the noise is not known, we provide an extension of our architecture, which we call RCGAN-U, that learns the noise model simultaneously while training the generator. We show experimentally on MNIST and CIFAR-10 datasets that both the approaches consistently improve upon baseline approaches, and RCGAN-U closely matches the performance of RCGAN.

Bias and Generalization in Deep Generative Models: An

Empirical Study

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #6

Shengjia Zhao · Hongyu Ren · Arianna Yuan · Jiaming Song · Noah Goodman · Stefano Ermon

In high dimensional settings, density estimation algorithms rely crucially on their inductive bias. Despite recent empirical success, the inductive bias of deep generative models is not well understood. In this paper we propose a framework to systematically investigate bias and generalization in deep generative models of images by probing the learning algorithm with carefully designed training datasets. By measuring properties of the learned distribution, we are able to find interesting patterns of generalization. We verify that these patterns are consistent across datasets, common models and architectures.

Probabilistic Neural Programmed Networks for Scene Generation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #7

Zhiwei Deng · Jiacheng Chen · YIFANG FU · Greg Mori

In this paper we address the text to scene image generation problem. Generative models that capture the variability in complicated scenes containing rich semantics is a grand goal of image generation. Complicated scene images contain rich visual elements, compositional visual concepts, and complicated relations between objects. Generative models, as an analysis-by-synthesis process, should encompass the following three core components: 1) the generation process that composes the scene; 2) what are the primitive visual elements and how are they composed; 3) the rendering of abstract concepts into their pixel-level realizations. We propose PNP-Net, a variational auto-encoder framework that addresses these three challenges: it flexibly composes images with a dynamic network structure, learns a set of distribution transformers that can compose distributions based on semantics, and decodes samples from these distributions into realistic images.

Step Size Matters in Deep Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #8

Kamil Nar · Shankar Sastry

Training a neural network with the gradient descent algorithm gives rise to a discrete-time nonlinear dynamical system. Consequently, behaviors that are typically observed in these systems emerge during training, such as convergence to an orbit but not to a fixed point or dependence of convergence on the initialization. Step size of the algorithm plays a critical role in these behaviors: it determines the subset of the local optima that the algorithm can converge to, and it specifies the magnitude of the oscillations if the algorithm converges to an orbit. To elucidate the effects of the step size on training of neural networks, we study the gradient descent algorithm as a discrete-time dynamical system, and by analyzing the Lyapunov stability of different solutions, we show the relationship between the step size of the algorithm and the solutions that can be obtained with this algorithm. The results provide an explanation for several phenomena observed in practice, including the deterioration in the training error with increased depth, the hardness of estimating linear mappings with large singular values, and the distinct performance of deep residual networks.

Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #9

Arthur Jacot-Guillarmod · Clement Hongler · Franck Gabriel

At initialization, artificial neural networks (ANNs) are equivalent to Gaussian processes in the infinite-width limit, thus connecting them to kernel methods. We prove that the evolution of an ANN during training can also be described by a kernel: during gradient descent on the parameters of an ANN, the network function (which maps input vectors to output vectors) follows the so-called kernel gradient associated with a new object, which we call the Neural Tangent Kernel (NTK). This kernel is central to describe the generalization features of ANNs. While the NTK is random at initialization and varies during training, in the infinite-width limit it converges to an explicit limiting kernel and stays constant during training. This makes it possible to study the training of ANNs in function space instead of parameter space. Convergence of the training can then be related to the positive-definiteness of the limiting NTK. We then focus on the setting of least-squares regression and show that in the infinite-width limit, the network function follows a linear differential equation during training. The convergence is fastest along the largest kernel principal components of the input data with respect to the NTK, hence suggesting a theoretical motivation for early stopping. Finally we study the NTK numerically, observe its behavior for wide networks, and compare it to the infinite-width limit.

How Does Batch Normalization Help Optimization?

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #10

Shibani Santurkar · Dimitris Tsipras · Andrew Ilyas · Aleksander Madry

Batch Normalization (BatchNorm) is a widely adopted technique that enables faster and more stable training of deep neural networks (DNNs). Despite its pervasiveness, the exact reasons for

BatchNorm's effectiveness are still poorly understood. The popular belief is that this effectiveness stems from controlling the change of the layers' input distributions during training to reduce the so-called "internal covariate shift". In this work, we demonstrate that such distributional stability of layer inputs has little to do with the success of BatchNorm. Instead, we uncover a more fundamental impact of BatchNorm on the training process: it makes the optimization landscape significantly smoother. This smoothness induces a more predictive and stable behavior of the gradients, allowing for faster training.

Towards Robust Detection of Adversarial Examples

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #11

Tianyu Pang · Chao Du · Yinpeng Dong · Jun Zhu

Although the recent progress is substantial, deep learning methods can be vulnerable to the maliciously generated adversarial examples. In this paper, we present a novel training procedure and a thresholding test strategy, towards robust detection of adversarial examples. In training, we propose to minimize the reverse cross-entropy (RCE), which encourages a deep network to learn latent representations that better distinguish adversarial examples from normal ones. In testing, we propose to use a thresholding strategy as the detector to filter out adversarial examples for reliable predictions. Our method is simple to implement using standard algorithms, with little extra training cost compared to the common cross-entropy minimization. We apply our method to defend various attacking methods on the widely used MNIST and CIFAR-10 datasets, and achieve significant improvements on robust predictions under all the threat models in the adversarial setting.

Training Neural Networks Using Features Replay

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #12

Zhouyuan Huo · Bin Gu · Heng Huang

Training a neural network using backpropagation algorithm requires passing error gradients sequentially through the network. The backward locking prevents us from updating network layers in parallel and fully leveraging the computing resources. Recently, there are several works trying to decouple and parallelize the backpropagation algorithm. However, all of them suffer from severe accuracy loss or memory explosion when the neural network is deep. To address these challenging issues, we propose a novel parallel-objective formulation for the objective function of the neural network. After that, we introduce features replay algorithm and prove that it is guaranteed to converge to critical points for the non-convex problem under certain conditions. Finally, we apply our method to training deep convolutional neural networks, and the experimental results show that the proposed method achieves {faster} convergence, {lower} memory consumption, and {better} generalization error than compared methods.

Faster Neural Networks Straight from JPEG

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #13

Lionel Gueguen · Alex Sergeev · Ben Kadlec · Rosanne Liu · Jason Yosinski

The simple, elegant approach of training convolutional neural networks (CNNs) directly from RGB pixels has enjoyed overwhelming empirical success. But can more performance be squeezed out of networks by using different input representations? In this paper we propose and explore a simple idea: train CNNs directly on the blockwise discrete cosine transform (DCT) coefficients computed and available in the middle of the JPEG codec. Intuitively, when processing JPEG images using CNNs, it seems unnecessary to decompress a blockwise frequency representation to an expanded pixel representation, shuffle it from CPU to GPU, and then process it with a CNN that will learn something similar to a transform back to frequency representation in its first layers. Why not skip both steps and feed the frequency domain into the network directly? In this paper we modify \libjpeg to produce DCT coefficients directly, modify a ResNet-50 network to accommodate the differently sized and strided input, and evaluate performance on ImageNet. We find networks that are both faster and more accurate, as well as networks with about the same accuracy but 1.77x faster than ResNet-50.

Hierarchical Graph Representation Learning with Differentiable Pooling

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #14

Zhitao Ying · Jiaxuan You · Christopher Morris · Xiang Ren · Will Hamilton · Jure Leskovec

Recently, graph neural networks (GNNs) have revolutionized the field of graph representation learning through effectively learned node embeddings, and achieved state-of-the-art results in tasks such as node classification and link prediction. However, current GNN methods are inherently flat and do not learn hierarchical representations of graphs---a limitation that is especially problematic for the task of graph classification, where the goal is to predict the label associated with an entire graph. Here we propose DiffPool, a differentiable graph pooling module that can generate hierarchical representations of graphs and can be combined with various graph neural network architectures in an end-to-end fashion. DiffPool learns a differentiable soft cluster assignment for nodes at each layer of a deep GNN, mapping nodes to a set of clusters, which then form the coarsened input for the next GNN layer. Our experimental results show that combining existing GNN methods with DiffPool yields an average improvement of 5-10% accuracy on graph classification benchmarks, compared to all existing pooling approaches, achieving a new state-of-the-art on four out of five benchmark datasets.

Bayesian Model-Agnostic Meta-Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #15

Jaesik Yoon · Taesup Kim · Ousmane Dia · Sungwoong Kim · Yoshua Bengio · Sungjin Ahn

Due to the inherent model uncertainty, learning to infer Bayesian posterior from a few-shot dataset is an important step towards robust meta-learning. In this paper, we propose a novel Bayesian model-agnostic meta-learning method. The proposed method combines efficient gradient-based meta-learning with nonparametric variational inference in a principled probabilistic framework. Unlike previous methods, during fast adaptation, the method is capable of learning complex uncertainty structure beyond a simple Gaussian approximation, and during meta-update, a novel Bayesian mechanism prevents meta-level overfitting. Remaining a gradient-based method, it is also the first Bayesian model-agnostic meta-learning method applicable to various tasks including reinforcement learning. Experiment results show the accuracy and robustness of the proposed method in sinusoidal regression, image classification, active learning, and reinforcement learning.

Evolved Policy Gradients

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #16

Rein Houthooft · Yuhua Chen · Phillip Isola · Bradly Stadie · Filip Wolski · OpenAI Jonathan Ho · Pieter Abbeel

We propose a metalearning approach for learning gradient-based reinforcement learning (RL) algorithms. The idea is to evolve a differentiable loss function, such that an agent, which optimizes its policy to minimize this loss, will achieve high rewards. The loss is parametrized via temporal convolutions over the agent's experience. Because this loss is highly flexible in its ability to take into account the agent's history, it enables fast task learning. Empirical results show that our evolved policy gradient algorithm (EPG) achieves faster learning on several randomized environments compared to an off-the-shelf policy gradient method. We also demonstrate that EPG's learned loss can generalize to out-of-distribution test time tasks, and exhibits qualitatively different behavior from other popular metalearning algorithms.

BourGAN: Generative Networks with Metric Embeddings

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #17

Chang Xiao · Peilin Zhong · Changxi Zheng

This paper addresses the mode collapse for generative adversarial networks (GANs). We view modes as a geometric structure of data distribution in a metric space. Under this geometric lens, we embed subsamples of the dataset from an arbitrary metric space into the L2 space, while preserving their pairwise distance distribution. Not only does this metric embedding determine the dimensionality of the latent space automatically, it also enables us to construct a mixture of Gaussians to draw latent space random vectors. We use the Gaussian mixture model in tandem with a simple augmentation of the objective function to train GANs. Every major step of our method is supported by theoretical analysis, and our experiments on real and synthetic data confirm that the generator is able to produce samples spreading over most of the modes while avoiding unwanted samples, outperforming several recent GAN variants on a number of metrics and offering new features.

Scalable End-to-End Autonomous Vehicle Testing via Rare-event Simulation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #18

Matthew O'Kelly · Aman Sinha · Hongseok Namkoong · Russ Tedrake · John Duchi

While recent developments in autonomous vehicle (AV) technology highlight substantial progress, we lack tools for rigorous and scalable testing. Real-world testing, the de facto evaluation environment, places the public in danger, and, due to the rare nature of accidents, will require billions of miles in order to statistically validate performance claims. We implement a simulation framework that can test an entire modern autonomous driving system, including, in particular, systems that employ deep-learning perception and control algorithms. Using adaptive importance-sampling methods to accelerate rare-event probability evaluation, we estimate the probability of an accident under a base distribution governing standard traffic behavior. We demonstrate our framework on a highway scenario, accelerating system evaluation by 2-20 times over naive Monte Carlo sampling methods and 10-300P times (where P is the number of processors) over real-world testing.

Generalisation of structural knowledge in the hippocampal-entorhinal system

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #19

James Whittington · Timothy Muller · Shirely Mark · Caswell Barry · Tim Behrens

A central problem to understanding intelligence is the concept of generalisation. This allows previously learnt structure to be exploited to solve tasks in novel situations differing in their particularities. We take inspiration from neuroscience, specifically the hippocampal-entorhinal system known to be important for generalisation. We propose that to generalise structural knowledge, the representations of the structure of the world, i.e. how entities in the world relate to each other, need to be separated from representations of the entities themselves. We show, under these principles, artificial neural networks embedded with hierarchy and fast Hebbian memory, can learn the statistics of memories and generalise structural knowledge. Spatial neuronal representations mirroring those found in the brain emerge, suggesting spatial cognition is an instance of more general organising principles. We further unify many entorhinal cell types as basis functions for constructing transition graphs, and show these representations effectively utilise memories. We experimentally support model assumptions, showing a preserved relationship between entorhinal grid and hippocampal place cells across environments.

Task-Driven Convolutional Recurrent Models of the Visual System

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #20

Aran Nayebi · Daniel Bear · Jonas Kubilius · Kohitij Kar · Surya Ganguli · David Sussillo · James J DiCarlo · Daniel Yamins

Feed-forward convolutional neural networks (CNNs) are currently state-of-the-art for object classification tasks such as ImageNet. Further, they are quantitatively accurate models of temporally-averaged responses of neurons in the primate brain's visual system. However, biological visual systems have two ubiquitous architectural features not shared with typical CNNs: local recurrence within cortical areas, and long-range feedback from downstream areas to upstream areas. Here we explored the role of recurrence in improving classification performance. We found that standard forms of recurrence (vanilla RNNs and LSTMs) do not perform well within deep CNNs on the ImageNet task. In contrast, novel cells that incorporated two structural features, bypassing and gating, were able to boost task accuracy substantially. We extended these design principles in an automated search over thousands of model architectures, which identified novel local recurrent cells and long-range feedback connections useful for object recognition. Moreover, these task-optimized ConvRNNs matched the dynamics of neural activity in the primate visual system better than feedforward networks, suggesting a role for the brain's recurrent connections in performing difficult visual behaviors.

Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #21

Kexin Yi · Jiajun Wu · Chuang Gan · Antonio Torralba · Pushmeet Kohli · Josh Tenenbaum

We marry two powerful ideas: deep representation learning for visual recognition and language understanding, and symbolic program execution for reasoning. Our neural-symbolic visual question answering (NS-VQA) system first recovers a structural scene representation from the image and a program trace from the question. It then executes the program on the scene representation to obtain an answer. Incorporating symbolic structure as prior knowledge offers three unique advantages. First, executing programs on a symbolic space is more robust to long program traces; our model can solve complex reasoning tasks better, achieving an accuracy of 99.8% on the CLEVR dataset. Second, the model is more data- and memory-efficient: it performs well after learning on a small number of training data; it can also encode an image into a compact representation, requiring less storage than existing methods for offline question answering. Third, symbolic program execution offers full transparency to the reasoning process; we are thus able to interpret and diagnose each execution step.

Extracting Relationships by Multi-Domain Matching

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #22

Yitong Li · michael Murias · geraldine Dawson · David Carlson

In many biological and medical contexts, we construct a large labeled corpus by aggregating many sources to use in target prediction tasks. Unfortunately, many of the sources may be irrelevant to our target task, so ignoring the structure of the dataset is detrimental. This work proposes a novel approach, the Multiple Domain Matching Network (MDMN), to exploit this structure. MDMN embeds all data into a shared feature space while learning which domains share strong statistical relationships. These relationships are often insightful in their own right, and they allow domains to share strength without interference from irrelevant data. This methodology builds on existing distribution-matching approaches by assuming that source domains are varied and outcomes multi-factorial. Therefore, each domain should only match a relevant subset. Theoretical analysis shows that the proposed approach can have a tighter generalization bound than existing multiple-domain adaptation approaches. Empirically, we show that the proposed methodology handles higher numbers of source domains (up to 21 empirically), and provides state-of-the-art performance on image, text, and multi-channel time series classification, including clinically relevant data of a novel treatment of Autism Spectrum Disorder.

Sparse Attentive Backtracking: Temporal Credit Assignment Through Reminding

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #23

Nan Rosemary Ke · Anirudh Goyal ALIAS PARTH GOYAL · Olexa Bilaniuk · Jonathan Binas · Michael Mozer · Chris Pal · Yoshua Bengio

Learning long-term dependencies in extended temporal sequences requires credit assignment to events far back in the past. The most common method for training recurrent neural networks, back-propagation through time (BPTT), requires credit information to be propagated backwards through every single step of the forward computation, potentially over thousands or millions of time steps. This becomes computationally expensive or even infeasible when used with long sequences. Importantly, biological brains are unlikely to perform such detailed reverse replay over very long sequences of internal states (consider days, months, or years.) However, humans are often reminded of past memories or mental states which are associated with the current mental state. We consider the hypothesis that such memory associations between past and present could be used for credit assignment through arbitrarily long sequences, propagating the credit assigned to the current state to the associated past state. Based on this principle, we study a novel algorithm which only back-propagates through a few of these temporal skip connections, realized by a learned attention mechanism that associates current states with relevant past states. We demonstrate in experiments that our method matches or outperforms regular BPTT and truncated BPTT in tasks involving particularly long-term dependencies, but without requiring the biologically implausible backward replay through the whole history of states. Additionally, we demonstrate that the proposed method transfers to longer sequences significantly better than LSTMs trained with BPTT and LSTMs trained with full self-attention.

Beauty-in-averageness and its contextual modulations: A Bayesian statistical account

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #24

Chaitanya Ryali · Angela J Yu

Understanding how humans perceive the likability of high-dimensional objects'' such as faces is an important problem in both cognitive science and AI/ML. Existing models generally assume these preferences to be fixed. However, psychologists have found human assessment of facial attractiveness to be context-dependent. Specifically, the classical Beauty-in-Averageness (BiA) effect, whereby a blended face is judged to be more attractive than the originals, is significantly diminished or reversed when the original faces are recognizable, or when the blend is mixed-race/mixed-gender and the attractiveness judgment is preceded by a race/gender categorization, respectively. This "Ugliness-in-Averageness" (UiA) effect has previously been explained via a qualitative disfluency account, which posits that the negative affect associated with the difficult race or gender categorization is inadvertently interpreted by the brain as a dislike for the face itself. In contrast, we hypothesize that human preference for an object is increased when it incurs lower encoding cost, in particular when its perceived {\it statistical typicality} is high, in consonance with Barlow's seminalefficient coding hypothesis.'' This statistical coding cost account explains both BiA, where

facial blends generally have higher likelihood than ``parent faces'', and UiA, when the preceding context or task restricts face representation to a task-relevant subset of features, thus redefining statistical typicality and encoding cost within that subspace. We use simulations to show that our model provides a parsimonious, statistically grounded, and quantitative account of both BiA and UiA. We validate our model using experimental data from a gender categorization task. We also propose a novel experiment, based on model predictions, that will be able to arbitrate between the disfluency account and our statistical coding cost account of attractiveness.

Stimulus domain transfer in recurrent models for large scale cortical population prediction on video

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #25

Fabian Sinz · Alexander Ecker · Paul Fahey · Edgar Walker · Erick Cobos · Emmanouil Froudarakis · Dimitri Yatsenko · Zachary Pitkow · Jacob Reimer · Andreas Tolias

To better understand the representations in visual cortex, we need to generate better predictions of neural activity in awake animals presented with their ecological input: natural video. Despite recent advances in models for static images, models for predicting responses to natural video are scarce and standard linear-nonlinear models perform poorly. We developed a new deep recurrent network architecture that predicts inferred spiking activity of thousands of mouse V1 neurons simultaneously recorded with two-photon microscopy, while accounting for confounding factors such as the animal's gaze position and brain state changes related to running state and pupil dilation. Powerful system identification models provide an opportunity to gain insight into cortical functions through in silico experiments that can subsequently be tested in the brain. However, in many cases this approach requires that the model is able to generalize to stimulus statistics that it was not trained on, such as band-limited noise and other parameterized stimuli. We investigated these domain transfer properties in our model and find that our model trained on natural images is able to correctly predict the orientation tuning of neurons in responses to artificial noise stimuli. Finally, we show that we can fully generalize from movies to noise and maintain high predictive performance on both stimulus domains by fine-tuning only the final layer's weights on a network otherwise trained on natural movies. The converse, however, is not true.

A probabilistic population code based on neural samples

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #26

Sabyasachi Shivkumar · Richard Lange · Ankani Chattoraj · Ralf Haefner

Sensory processing is often characterized as implementing probabilistic inference: networks of neurons compute posterior beliefs over unobserved causes given the sensory inputs. How these beliefs are computed and represented by neural responses is much-debated (Fiser et al. 2010, Pouget et al. 2013). A central debate concerns the question of whether neural responses represent samples of latent variables (Hoyer & Hyvarinnen 2003) or parameters of their distributions (Ma et al. 2006) with efforts being made to distinguish between them (Grabska-Barwinska et al. 2013). A separate debate addresses the question of whether neural responses are proportionally related to the encoded probabilities (Barlow 1969), or proportional to the logarithm of those probabilities (Jazayeri & Movshon 2006, Ma et al. 2006, Beck et al. 2012). Here, we show that these alternatives -- contrary to common assumptions -- are not mutually exclusive and that the very same system can be compatible with all of them. As a central analytical result, we show that modeling neural responses in area V1 as samples from a posterior distribution over latents in a linear Gaussian model of the image implies that those neural responses form a linear Probabilistic Population Code (PPC, Ma et al. 2006). In particular, the posterior distribution over some experimenter-defined variable like "orientation" is part of the exponential family with sufficient statistics that are linear in the neural sampling-based firing rates.

---

cpSGD: Communication-efficient and differentially-private distributed SGD

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #27

Naman Agarwal · Ananda Theertha Suresh · Felix Xinnan Yu · Sanjiv Kumar · Brendan McMahan

Distributed stochastic gradient descent is an important subroutine in distributed learning. A setting of particular interest is when the clients are mobile devices, where two important concerns are communication efficiency and the privacy of the clients. Several recent works have focused on reducing the communication cost or introducing privacy guarantees, but none of the proposed communication efficient methods are known to be privacy preserving and none of the known privacy mechanisms are known to be communication efficient. To this end, we study algorithms that achieve both communication efficiency and differential privacy. For $d$ variables and $n \approx d$ clients, the proposed method uses $\cO(\log \log(nd))$ bits of communication per client per coordinate and ensures constant privacy.

We also improve previous analysis of the \emph{Binomial mechanism} showing that it achieves nearly the same utility as the Gaussian mechanism, while requiring fewer representation bits, which can be of independent interest.

Bounded-Loss Private Prediction Markets

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #28

Rafael Frongillo · Bo Waggoner

Prior work has investigated variations of prediction markets that preserve participants' (differential) privacy, which formed the basis of useful mechanisms for purchasing data for machine learning objectives. Such markets required potentially unlimited financial subsidy, however, making them impractical. In this work, we design an adaptively-growing prediction market with a bounded financial subsidy, while achieving privacy, incentives to produce accurate predictions, and precision in the sense that market prices are not heavily impacted by the added privacy-preserving noise. We briefly discuss how our mechanism can extend to the data-purchasing setting, and its relationship to traditional learning algorithms.

Ex ante coordination and collusion in zero-sum multi-player extensive-form games

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #29

Gabriele Farina · Andrea Celli · Nicola Gatti · Tuomas Sandholm

Recent milestones in equilibrium computation, such as the success of Libratus, show that it is possible to compute strong solutions to two-player zero-sum games in theory and practice. This is not the case for games with more than two players, which remain one of the main open challenges in computational game theory. This paper focuses on zero-sum games where a team of players faces an opponent, as is the case, for example, in Bridge, collusion in poker, and many non-recreational applications such as war, where the colluders do not have time or means of communicating during battle, collusion in bidding, where communication during the auction is illegal, and coordinated swindling in public. The possibility for the team members to communicate before game play—that is, coordinate their strategies ex ante—makes the use of behavioral strategies unsatisfactory. The reasons for this are closely related to the fact that the team can be represented as a single player with imperfect recall. We propose a new game representation, the realization form, that generalizes the sequence form but can also be applied to imperfect-recall games. Then, we use it to derive an auxiliary game that is equivalent to the original one. It provides a sound way to map the problem of finding an optimal ex-ante-correlated strategy for the team to the well-understood Nash equilibrium-finding problem in a (larger) two-player zero-sum perfect-recall game. By reasoning over the auxiliary game, we devise an anytime algorithm, fictitious team-play, that is guaranteed to converge to an optimal coordinated strategy for the team against an optimal opponent, and that is dramatically faster than the prior state-of-the-art algorithm for this problem.

Model-Agnostic Private Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #30

Raef Bassily · Abhradeep Guha Thakurta · Om Dipakbhai Thakkar

We design differentially private learning algorithms that are agnostic to the learning model assuming access to limited amount of unlabeled public data. First, we give a new differentially private algorithm for answering a sequence of m online classification queries (given by a sequence of m unlabeled public feature vectors) based on a private training set. Our private algorithm follows the paradigm of subsample-and-aggregate, in which any generic non-private learner is trained on disjoint subsets of the private training set, then for each classification query, the votes of the resulting classifiers ensemble are aggregated in a differentially private fashion. Our private aggregation is based on a novel combination of distance-to-instability framework [Smith & Thakurta 2013] and the sparse-vector technique [Dwork et al. 2009, Hardt & Talwar 2010]. We show that our algorithm makes a conservative use of the privacy budget. In particular, if the underlying non-private learner yields classification error at most $\alpha\in (0, 1)$, then our construction answers more queries, by at least a factor of $1/\alpha$ in some cases, than what is implied by a straightforward application of the advanced composition theorem for differential privacy. Next, we apply the knowledge transfer technique to construct a private learner that outputs a classifier, which can be used to answer unlimited number of queries. In the PAC model, we analyze our construction and prove upper bounds on the sample complexity for both the realizable and the non-realizable cases. As in non-private sample complexity, our bounds are completely characterized by the VC dimension of the concept class.

Adversarially Robust Generalization Requires More Data

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #31

Ludwig Schmidt · Shibani Santurkar · Dimitris Tsipras · Kunal Talwar · Aleksander Madry

Machine learning models are often susceptible to adversarial perturbations of their inputs. Even small perturbations can cause state-of-the-art classifiers with high "standard" accuracy to produce an incorrect prediction with high confidence. To better understand this phenomenon, we study adversarially robust learning from the viewpoint of generalization. We show that already in a simple natural data model, the sample complexity of robust learning can be significantly larger than that of "standard" learning. This gap is information theoretic and holds irrespective of the training algorithm or the model family. We complement our theoretical results with experiments on popular image classification datasets and show that a similar gap exists here as well. We postulate that the difficulty of training robust classifiers stems, at least partially, from this inherently larger sample complexity.

Probabilistic Pose Graph Optimization via Bingham Distributions and Tempered Geodesic MCMC

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #32

Tolga Birdal · Umut Simsekli · Mustafa Onur Eken · Slobodan Ilic

We introduce Tempered Geodesic Markov Chain Monte Carlo (TG-MCMC) algorithm for initializing pose graph optimization problems, arising in various scenarios such as SFM (structure from motion) or SLAM (simultaneous localization and mapping). TG-MCMC is first of its kind as it unites global non-convex optimization on the spherical manifold of quaternions with posterior sampling, in order to provide both reliable initial poses and uncertainty estimates that are informative about the quality of solutions. We devise theoretical convergence guarantees and extensively evaluate our method on synthetic and real benchmarks. Besides its elegance in formulation and theory, we show that our method is robust to missing data, noise and the estimated uncertainties capture intuitive properties of the data.

A Statistical Recurrent Model on the Manifold of Symmetric Positive Definite Matrices

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #33

Rudrasis Chakraborty · Chun-Hao Yang · Xingjian Zhen · Monami Banerjee · Derek Archer · David Vaillancourt · Vikas Singh · Baba Vemuri

In a number of disciplines, the data (e.g., graphs, manifolds) to be analyzed are non-Euclidean in nature. Geometric deep learning corresponds to techniques that generalize deep neural network models to such non-Euclidean spaces. Several recent papers have shown how convolutional neural networks (CNNs) can be extended to learn with graph-based data. In this work, we study the setting where the data (or measurements) are ordered, longitudinal or temporal in nature and live on a Riemannian manifold -- this setting is common in a variety of problems in statistical machine learning, vision and medical imaging. We show how recurrent statistical recurrent network models can be defined in such spaces. We give an efficient algorithm and conduct a rigorous analysis of its statistical properties. We perform extensive numerical experiments demonstrating competitive performance with state of the art methods but with significantly less number of parameters. We also show applications to a statistical analysis task in brain imaging, a regime where deep neural network models have only been utilized in limited ways.

Designing by Training: Acceleration Neural Network for Fast High-Dimensional Convolution

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #34

Longquan Dai · Liang Tang · Yuan Xie · Jinhui Tang

The high-dimensional convolution is widely used in various disciplines but has a serious performance problem due to its high computational complexity. Over the decades, people took a handmade approach to design fast algorithms for the Gaussian convolution. Recently, requirements for various non-Gaussian convolutions have emerged and are continuously getting higher. However, the handmade acceleration approach is no longer feasible for so many different convolutions since it is a time-consuming and painstaking job. Instead, we propose an Acceleration Network (AccNet) which turns the work of designing new fast algorithms to training the AccNet. This is done by: 1, interpreting splatting, blurring, slicing operations as convolutions; 2, turning these convolutions to $g$CP layers to build AccNet. After training, the activation function $g$ together with AccNet weights automatically define the new splatting, blurring and slicing operations. Experiments demonstrate AccNet is able to design acceleration algorithms for a ton of convolutions including Gaussian/non-Gaussian convolutions and produce state-of-the-art results.

Greedy Hash: Towards Fast Optimization for Accurate Hash Coding in CNN

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #35

Shupeng Su · Chao Zhang · Kai Han · Yonghong Tian

To convert the input into binary code, hashing algorithm has been widely used for approximate nearest neighbor search on large-scale image sets due to its computation and storage efficiency. Deep hashing further improves the retrieval quality by combining the hash coding with deep neural network. However, a major difficulty in deep hashing lies in the discrete constraints imposed on the network output, which generally makes the optimization NP hard. In this work, we adopt the greedy principle to tackle this NP hard problem by iteratively updating the network toward the probable optimal discrete solution in each iteration. A hash coding layer is designed to implement our approach which strictly uses the sign function in forward propagation to maintain the discrete constraints, while in back propagation the gradients are transmitted intactly to the front layer to avoid the vanishing gradients. In addition to the theoretical derivation, we provide a new perspective to visualize and understand the effectiveness and efficiency of our algorithm. Experiments on benchmark datasets show that our scheme outperforms state-of-the-art hashing methods in both supervised and unsupervised tasks.

Generative Probabilistic Novelty Detection with Adversarial Autoencoders

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #36

Stanislav Pidhorskyi · Ranya Almohsen · Gianfranco Doretto

Novelty detection is the problem of identifying whether a new data point is considered to be an inlier or an outlier. We assume that training data is available to describe only the inlier distribution. Recent approaches primarily leverage deep encoder-decoder network architectures to compute a reconstruction error that is used to either compute a novelty score or to train a one-class classifier. While we too leverage a novel network of that kind, we take a probabilistic approach and effectively compute how likely it is that a sample was generated by the inlier distribution. We achieve this with two main contributions. First, we make the computation of the novelty probability feasible because we linearize the parameterized manifold capturing the underlying structure of the inlier distribution, and show how the probability factorizes and can be computed with respect to local coordinates of the manifold tangent space. Second, we improve the training of the autoencoder network. An extensive set of results show that the approach achieves state-of-the-art performance on several benchmark datasets.

Joint Sub-bands Learning with Clique Structures for Wavelet Domain Super-Resolution

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #37

Zhisheng Zhong · Tiancheng Shen · Yibo Yang · Zhouchen Lin · Chao Zhang

Convolutional neural networks (CNNs) have recently achieved great success in single-image super-resolution (SISR). However, these methods tend to produce over-smoothed outputs and miss some textural details. To solve these problems, we propose the Super-Resolution CliqueNet (SRCliqueNet) to reconstruct the high resolution (HR) image with better textural details in the wavelet domain. The proposed SRCliqueNet firstly extracts a set of feature maps from the low resolution (LR) image by the clique blocks group. Then we send the set of feature maps to the clique up-sampling module to reconstruct the HR image. The clique up-sampling module consists of four sub-nets which predict the high resolution wavelet coefficients of four sub-bands. Since we consider the edge feature properties of four sub-bands, the four sub-nets are connected to the others so that they can learn the coefficients of four sub-bands jointly. Finally we apply inverse discrete wavelet transform (IDWT) to the output of four sub-nets at the end of the clique up-sampling module to increase the resolution and reconstruct the HR image. Extensive quantitative and qualitative experiments on benchmark datasets show that our method achieves superior performance over the state-of-the-art methods.

Compact Generalized Non-local Network

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #38

Kaiyu Yue · Ming Sun · Yuchen Yuan · Feng Zhou · Errui Ding · Fuxin Xu

The non-local module is designed for capturing long-range spatio-temporal dependencies in images and videos. Although having shown excellent performance, it lacks the mechanism to model the interactions between positions across channels, which are of vital importance in recognizing fine-grained objects and actions. To address this limitation, we generalize the non-local module and take the correlations between the positions of any two channels into account. This extension utilizes the compact representation for multiple kernel functions with Taylor expansion that makes the generalized non-local module in a fast and low-complexity computation flow. Moreover, we implement our generalized non-local method within channel groups to ease the optimization. Experimental results illustrate the clear-cut improvements and practical applicability of the generalized non-local module on both fine-grained object recognition and video classification. Code is available at: https://github.com/KaiyuYue/cgnl-network.pytorch.

BinGAN: Learning Compact Binary Descriptors with a Regularized GAN

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #39

Maciej Zieba · Piotr Semberecki · Tarek El-Gaaly · Tomasz Trzcinski

In this paper, we propose a novel regularization method for Generative Adversarial Networks that allows the model to learn discriminative yet compact binary representations of image patches (image descriptors). We exploit the dimensionality reduction that takes place in the intermediate layers of the discriminator network and train the binarized penultimate layer's low-dimensional representation to mimic the distribution of the higher-dimensional preceding layers. To achieve this, we introduce two loss terms that aim at: (i) reducing the correlation between the dimensions of the binarized penultimate layer's low-dimensional representation (i.e. maximizing joint entropy) and (ii) propagating the relations between the dimensions in the high-dimensional space to the low-dimensional space. We evaluate the resulting binary image descriptors on two challenging applications, image matching and retrieval, where they achieve state-of-the-art results.

RenderNet: A deep convolutional network for differentiable rendering from 3D shapes

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #40

Thu H Nguyen-Phuoc · Chuan Li · Stephen Balaban · Yongliang Yang

Traditional computer graphics rendering pipelines are designed for procedurally generating 2D images from 3D shapes with high performance. The nondifferentiability due to discrete operations (such as visibility computation) makes it hard to explicitly correlate rendering parameters and the resulting image, posing a significant challenge for inverse rendering tasks. Recent work on differentiable rendering achieves differentiability either by designing surrogate gradients for non-differentiable operations or via an approximate but differentiable renderer. These methods, however, are still limited when it comes to handling occlusion, and restricted to particular rendering effects. We present RenderNet, a differentiable rendering convolutional network with a novel projection unit that can render 2D images from 3D shapes. Spatial occlusion and shading calculation are automatically encoded in the network. Our experiments show that RenderNet can successfully learn to implement different shaders, and can be used in inverse rendering tasks to estimate shape, pose, lighting and texture from a single image.

LF-Net: Learning Local Features from Images

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #41

Yuki Ono · Eduard Trulls · Pascal Fua · Kwang Moo Yi

We present a novel deep architecture and a training strategy to learn a local feature pipeline from scratch, using collections of images without the need for human supervision. To do so we exploit depth and relative camera pose cues to create a virtual target that the network should achieve on one image, provided the outputs of the network for the other image. While this process is inherently non-differentiable, we show that we can optimize the network in a two-branch setup by confining it to one branch, while preserving differentiability in the other. We train our method on both indoor and outdoor datasets, with depth data from 3D sensors for the former, and depth estimates from an off-the-shelf Structure-from-Motion solution for the latter. Our models outperform the state of the art on sparse feature matching on both datasets, while running at 60+ fps for QVGA images.

Unsupervised Learning of Shape and Pose with Differentiable Point Clouds

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #42

Eldar Insafutdinov · Alexey Dosovitskiy

We address the problem of learning accurate 3D shape and camera pose from a collection of unlabeled category-specific images. We train a convolutional network to predict both the shape and the pose from a single image by minimizing the reprojection error: given several views of an object, the projections of the predicted shapes to the predicted camera poses should match the provided views. To deal with pose ambiguity, we introduce an ensemble of pose predictors which we then distill to a single "student" model. To allow for efficient learning of high-fidelity shapes, we represent the shapes by point clouds and devise a formulation allowing for differentiable projection of these. Our experiments show that the distilled ensemble of pose predictors learns to estimate the pose accurately, while the point cloud representation allows to predict detailed shape models.

Modelling and unsupervised learning of symmetric deformable object categories

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #43

James Thewlis · Hakan Bilen · Andrea Vedaldi

We propose a new approach to model and learn, without manual supervision, the symmetries of natural objects, such as faces or flowers, given only images as input. It is well known that objects that have a symmetric structure do not usually result in symmetric images due to articulation and perspective effects. This is often tackled by seeking the intrinsic symmetries of the underlying 3D shape, which is very difficult to do when the latter cannot be recovered reliably from data. We show that, if only raw images are given, it is possible to look instead for symmetries in the space of object deformations. We can then learn symmetries from an unstructured collection of images of the object as an extension of the recently-introduced object frame representation, modified so that object symmetries reduce to the obvious symmetry groups in the normalized space. We also show that our formulation provides an explanation of the ambiguities that arise in recovering the pose of symmetric objects from their shape or images and we provide a way of discounting such ambiguities in learning.

Multi-View Silhouette and Depth Decomposition for High Resolution 3D Object Representation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #44

Edward Smith · Scott Fujimoto · David Meger

We consider the problem of scaling deep generative shape models to high-resolution. Drawing motivation from the canonical view representation of objects, we introduce a novel method for the fast up-sampling of 3D objects in voxel space through networks that perform super-resolution on the six orthographic depth projections. This allows us to generate high-resolution objects with more efficient scaling than methods which work directly in 3D. We decompose the problem of 2D depth super-resolution into silhouette and depth prediction to capture both structure and fine detail. This allows our method to generate sharp edges more easily than an individual network. We evaluate our work on multiple experiments concerning high-resolution 3D objects, and show our system is capable of accurately predicting novel objects at resolutions as large as 512x512x512 -- the highest resolution reported for this task. We achieve state-of-the-art performance on 3D object reconstruction from RGB images on the ShapeNet dataset, and further demonstrate the first effective 3D super-resolution method.

Image Inpainting via Generative Multi-column Convolutional Neural Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #45

Yi Wang · Xin Tao · Xiaojuan Qi · Xiaoyong Shen · Jiaya Jia

In this paper, we propose a generative multi-column network for image inpainting. This network synthesizes different image components in a parallel manner within one stage. To better characterize global structures, we design a confidence-driven reconstruction loss while an implicit diversified MRF regularization is adopted to enhance local details. The multi-column network combined with the reconstruction and MRF loss propagates local and global information derived from context to the target inpainting regions. Extensive experiments on challenging street view, face, natural objects and scenes manifest that our method produces visual compelling results even without previously common post-processing.

Beyond Grids: Learning Graph Representations for Visual Recognition

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #46

Yin Li · Abhinav Gupta

We propose learning graph representations from 2D feature maps for visual recognition. Our method draws inspiration from region based recognition, and learns to transform a 2D image into a graph structure. The vertices of the graph define clusters of pixels ("regions"), and the edges measure the similarity between these clusters in a feature space. Our method further learns to propagate information across all vertices on the graph, and is able to project the learned graph representation back into 2D grids. Our graph representation facilitates reasoning beyond regular grids and can capture long range dependencies among regions. We demonstrate that our model can be trained from end-to-end, and is easily integrated into existing networks. Finally, we evaluate our method on three challenging recognition tasks: semantic segmentation, object detection and object instance segmentation. For all tasks, our method outperforms state-of-the-art methods.

Foreground Clustering for Joint Segmentation and Localization in Videos and Images

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #47

Abhishek Sharma

This paper presents a novel framework in which video/image segmentation and localization are cast into a single optimization problem that integrates information from low level appearance cues with that of high level localization cues in a very weakly supervised manner. The proposed framework leverages two representations at different levels, exploits the spatial relationship between bounding boxes and superpixels as linear constraints and simultaneously discriminates between foreground and background at bounding box and superpixel level. Different from previous approaches that mainly rely on discriminative clustering, we incorporate a foreground model that minimizes the histogram difference of an object across all image frames. Exploiting the geometric relation between the superpixels and bounding boxes enables the transfer of segmentation cues to improve localization output and vice-versa. Inclusion of the foreground model generalizes our discriminative framework to video data where the background tends to be similar and thus, not discriminative. We demonstrate the effectiveness of our unified framework on the YouTube Object video dataset, Internet Object Discovery dataset and Pascal VOC 2007.

LinkNet: Relational Embedding for Scene Graph

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #48

Sanghyun Woo · Dahun Kim · Donghyeon Cho · In So Kweon

Objects and their relationships are critical contents for image understanding. A scene graph provides a structured description that captures these properties of an image. However, reasoning about the relationships between objects is very challenging and only a few recent works have attempted to solve the problem of generating a scene graph from an image. In this paper, we present a novel method that improves scene graph generation by explicitly modeling inter-dependency among the entire object instances. We design a simple and effective relational embedding module that enables our model to jointly represent connections among all related objects, rather than focus on an object in isolation. Our novel method significantly benefits two main parts of the scene graph generation task: object classification and relationship classification. Using it on top of a basic Faster R-CNN, our model achieves state-of-the-art results on the Visual Genome benchmark. We further push the performance by introducing global context encoding module and geometrical layout encoding module. We validate our final model, LinkNet, through extensive ablation studies, demonstrating its efficacy in scene graph generation.

Context-aware Synthesis and Placement of Object Instances

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #49

Donghoon Lee · Ming-Yu Liu · Ming-Hsuan Yang · Sifei Liu · Jinwei Gu · Jan Kautz

Learning to insert an object instance into an image in a semantically coherent manner is a challenging and interesting problem. Solving it requires (a) determining a location to place an object in the scene and (b) determining its appearance at the location. Such an object insertion model can potentially facilitate numerous image editing and scene parsing applications. In this paper, we propose an end-to-end trainable neural network for the task of inserting an object instance mask of a specified class into the semantic label map of an image. Our network consists of two generative modules where one determines where the inserted object mask should be (i.e., location and scale) and the other determines what the object mask shape (and pose) should look like. The two modules are connected together via a spatial transformation network and jointly trained. We devise a learning procedure that leverage both supervised and unsupervised data and show our model can insert an object at diverse locations with various appearances. We conduct extensive experimental validations with comparisons to strong baselines to verify the effectiveness of the proposed network.

Geometry-Aware Recurrent Neural Networks for Active Visual Recognition

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #50

Ricson Cheng · Ziyan Wang · Katerina Fragkiadaki

We present recurrent geometry-aware neural networks that integrate visual in- formation across multiple views of a scene into 3D latent feature tensors, while maintaining an one-to-one mapping between 3D physical locations in the world scene and latent feature locations. Object detection, object segmentation, and 3D reconstruction is then carried out directly using the constructed 3D feature memory, as opposed to any of the input 2D images. The proposed models are equipped with differentiable egomotion-aware feature warping and (learned) depth-aware unprojection operations to achieve geometrically consistent mapping between the features in the input frame and the constructed latent model of the scene. We empirically show the proposed model generalizes much better than geometry- unaware LSTM/GRU networks, especially under the presence of multiple objects and cross-object occlusions. Combined with active view selection policies, our model learns to select informative viewpoints to integrate information from by “undoing" cross-object occlusions, seamlessly combining geometry with learning from experience.

See and Think: Disentangling Semantic Scene Completion

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #51

Shice Liu · YU HU · Yiming Zeng · Qiankun Tang · Beibei Jin · Yinhe Han · Xiaowei Li

Semantic scene completion predicts volumetric occupancy and object category of a 3D scene, which helps intelligent agents to understand and interact with the surroundings. In this work, we propose a disentangled framework, sequentially carrying out 2D semantic segmentation, 2D-3D reprojection and 3D semantic scene completion. This three-stage framework has three advantages: (1) explicit semantic segmentation significantly boosts performance; (2) flexible fusion ways of sensor data bring good extensibility; (3) progress in any subtask will promote the holistic performance. Experimental results show that regardless of inputing a single depth or RGB-D, our framework can generate high-quality semantic scene completion, and outperforms state-of-the-art approaches on both synthetic and real datasets.

Active Matting

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #52

Xin Yang · Ke Xu · Shaozhe Chen · Shengfeng He · Baocai Yin Yin · Rynson Lau

Image matting is an ill-posed problem. It requires a user input trimap or some strokes to obtain an alpha matte of the foreground object. A fine user input is essential to obtain a good result, which is either time consuming or suitable for experienced users who know where to place the strokes. In this paper, we explore the intrinsic relationship between the user input and the matting algorithm to address the problem of where and when the user should provide the input. Our aim is to discover the most informative sequence of regions for user input in order to produce a good alpha matte with minimum labeling efforts. To this end, we propose an active matting method with recurrent reinforcement learning. The proposed framework involves human in the loop by sequentially detecting informative regions for trivial human judgement. Comparing to traditional matting algorithms, the proposed framework requires much less efforts, and can produce satisfactory results with just 10 regions. Through extensive experiments, we show that the proposed model reduces user efforts significantly and achieves comparable performance to dense trimaps in a user-friendly manner. We further show that the learned informative knowledge can be generalized across different matting algorithms.

A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #53

Alexander H. Liu · Yen-Cheng Liu · Yu-Ying Yeh · Yu-Chiang Frank Wang

We present a novel and unified deep learning framework which is capable of learning domain-invariant representation from data across multiple domains. Realized by adversarial training with additional ability to exploit domain-specific information, the proposed network is able to perform continuous cross-domain image translation and manipulation, and produces desirable output images accordingly. In addition, the resulting feature representation exhibits superior performance of unsupervised domain adaptation, which also verifies the effectiveness of the proposed model in learning disentangled features for describing cross-domain data.

Turbo Learning for CaptionBot and DrawingBot

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #54

Qiuyuan Huang · Pengchuan Zhang · Dapeng Wu · Lei Zhang

We study in this paper the problems of both image captioning and text-to-image generation, and present a novel turbo learning approach to jointly training an image-to-text generator (a.k.a. CaptionBot) and a text-to-image generator (a.k.a. DrawingBot). The key idea behind the joint training is that image-to-text generation and text-to-image generation as dual problems can form a closed loop to provide informative feedback to each other. Based on such feedback, we introduce a new loss metric by comparing the original input with the output produced by the closed loop. In addition to the old loss metrics used in CaptionBot and DrawingBot, this extra loss metric makes the jointly trained CaptionBot and DrawingBot better than the separately trained CaptionBot and DrawingBot. Furthermore, the turbo-learning approach enables semi-supervised learning since the closed loop can provide peudo-labels for unlabeled samples. Experimental results on the COCO dataset demonstrate that the proposed turbo learning can significantly improve the performance of both CaptionBot and DrawingBot by a large margin.

Dialog-based Interactive Image Retrieval

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #55

Xiaoxiao Guo · Hui Wu · Yu Cheng · Steven Rennie · Gerald Tesauro · Rogerio Feris

Existing methods for interactive image retrieval have demonstrated the merit of integrating user feedback, improving retrieval results. However, most current systems rely on restricted forms of user feedback, such as binary relevance responses, or feedback based on a fixed set of relative attributes, which limits their impact. In this paper, we introduce a new approach to interactive image search that enables users to provide feedback via natural language, allowing for more natural and effective interaction. We formulate the task of dialog-based interactive image retrieval as a reinforcement learning problem, and reward the dialog system for improving the rank of the target image during each dialog turn. To mitigate the cumbersome and costly process of collecting human-machine conversations as the dialog system learns, we train our system with a user simulator, which is itself trained to describe the differences between target and candidate images. The efficacy of our approach is demonstrated in a footwear retrieval application. Experiments on both simulated and real-world data show that 1) our proposed learning framework achieves better accuracy than other supervised and reinforcement learning baselines and 2) user feedback based on natural language rather than pre-specified attributes leads to more effective retrieval results, and a more natural and expressive communication interface.

Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #56

Yuan Li · Xiaodan Liang · Zhiting Hu · Eric Xing

Generating long and coherent reports to describe medical images poses challenges to bridging visual patterns with informative human linguistic descriptions. We propose a novel Hybrid Retrieval-Generation Reinforced Agent (HRGR-Agent) which reconciles traditional retrieval-based approaches populated with human prior knowledge, with modern learning-based approaches to achieve structured, robust, and diverse report generation. HRGR-Agent employs a hierarchical decision-making procedure. For each sentence, a high-level retrieval policy module chooses to either retrieve a template sentence from an off-the-shelf template database, or invoke a low-level generation module to generate a new sentence. HRGR-Agent is updated via reinforcement learning, guided by sentence-level and word-level rewards. Experiments show that our approach achieves the state-of-the-art results on two medical report datasets, generating well-balanced structured sentences with robust coverage of heterogeneous medical report contents. In addition, our model achieves the highest detection precision of medical abnormality terminologies, and improved human evaluation performance.

Sequential Context Encoding for Duplicate Removal

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #57

Lu Qi · Shu Liu · Jianping Shi · Jiaya Jia

Duplicate removal is a critical step to accomplish a reasonable amount of predictions in prevalent proposal-based object detection frameworks. Albeit simple and effective, most previous algorithms utilized a greedy process without making sufficient use of properties of input data. In this work, we design a new two-stage framework to effectively select the appropriate proposal candidate for each object. The first stage suppresses most of easy negative object proposals, while the second stage selects true positives in the reduced proposal set. These two stages share the same network structure, an encoder and a decoder formed as recurrent neural networks (RNN) with global attention and context gate. The encoder scans proposal candidates in a sequential manner to capture the global context information, which is then fed to the decoder to extract optimal proposals. In our extensive experiments, the proposed method outperforms other alternatives by a large margin.

Hybrid Knowledge Routed Modules for Large-scale Object Detection

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #58

ChenHan Jiang · Hang Xu · Xiaodan Liang · Liang Lin

Abstract The dominant object detection approaches treat the recognition of each region separately and overlook crucial semantic correlations between objects in one scene. This paradigm leads to substantial performance drop when facing heavy long-tail problems, where very few samples are available for rare classes and plenty of confusing categories exists. We exploit diverse human commonsense knowledge for reasoning over large-scale object categories and reaching semantic coherency within one image. Particularly, we present Hybrid Knowledge Routed Modules (HKRM) that incorporates the reasoning routed by two kinds of knowledge forms: an explicit knowledge module for structured constraints that are summarized with linguistic knowledge (e.g. shared attributes, relationships) about concepts; and an implicit knowledge module that depicts some implicit constraints (e.g. common spatial layouts). By functioning over a region-to-region graph, both modules can be individualized and adapted to coordinate with visual patterns in each image, guided by specific knowledge forms. HKRM are light-weight, general-purpose and extensible by easily incorporating multiple knowledge to endow any detection networks the ability of global semantic reasoning. Experiments on large-scale object detection benchmarks show HKRM obtains around 34.5% improvement on VisualGenome (1000 categories) and 30.4% on ADE in terms of mAP.

SNIPER: Efficient Multi-Scale Training

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #59

Bharat Singh · Mahyar Najibi · Larry Davis

We present SNIPER, an algorithm for performing efficient multi-scale training in instance level visual recognition tasks. Instead of processing every pixel in an image pyramid, SNIPER processes context regions around ground-truth instances (referred to as chips) at the appropriate scale. For background sampling, these context-regions are generated using proposals extracted from a region proposal network trained with a short learning schedule. Hence, the number of chips generated per image during training adaptively changes based on the scene complexity. SNIPER only processes 30% more pixels compared to the commonly used single scale training at 800x1333 pixels on the COCO dataset. But, it also observes samples from extreme resolutions of the image pyramid, like 1400x2000 pixels. As SNIPER operates on resampled low resolution chips (512x512 pixels), it can have a batch size as large as 20 on a single GPU even with a ResNet-101 backbone. Therefore it can benefit from batch-normalization during training without the need for synchronizing batch-normalization statistics across GPUs. SNIPER brings training of instance level recognition tasks like object detection closer to the protocol for image classification and suggests that the commonly accepted guideline that it is important to train on high resolution images for instance level visual recognition tasks might not be correct. Our implementation based on Faster-RCNN with a ResNet-101 backbone obtains an mAP of 47.6% on the COCO dataset for bounding box detection and can process 5 images per second during inference with a single GPU. Code is available at https://github.com/MahyarNajibi/SNIPER/ .

Revisiting Multi-Task Learning with ROCK: a Deep Residual Auxiliary Block for Visual Detection

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #60

Taylor Mordan · Nicolas THOME · Gilles Henaff · Matthieu Cord

Multi-Task Learning (MTL) is appealing for deep learning regularization. In this paper, we tackle a specific MTL context denoted as primary MTL, where the ultimate goal is to improve the performance of a given primary task by leveraging several other auxiliary tasks. Our main methodological contribution is to introduce ROCK, a new generic multi-modal fusion block for deep learning tailored to the primary MTL context. ROCK architecture is based on a residual connection, which makes forward prediction explicitly impacted by the intermediate auxiliary representations. The auxiliary predictor's architecture is also specifically designed to our primary MTL context, by incorporating intensive pooling operators for maximizing complementarity of intermediate representations. Extensive experiments on NYUv2 dataset (object detection with scene classification, depth prediction, and surface normal estimation as auxiliary tasks) validate the relevance of the approach and its superiority to flat MTL approaches. Our method outperforms state-of-the-art object detection models on NYUv2 by a large margin, and is also able to handle large-scale heterogeneous inputs (real and synthetic images) with missing annotation modalities.

MetaAnchor: Learning to Detect Objects with Customized Anchors

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #61

Tong Yang · Xiangyu Zhang · Zeming Li · Wenqiang Zhang · Jian Sun

We propose a novel and flexible anchor mechanism named MetaAnchor for object detection frameworks. Unlike many previous detectors model anchors via a predefined manner, in MetaAnchor anchor functions could be dynamically generated from the arbitrary customized prior boxes. Taking advantage of weight prediction, MetaAnchor is able to work with most of the anchor-based object detection systems such as RetinaNet. Compared with the predefined anchor scheme, we empirically find that MetaAnchor is more robust to anchor settings and bounding box distributions; in addition, it also shows the potential on the transfer task. Our experiment on COCO detection task shows MetaAnchor consistently outperforms the counterparts in various scenarios.

Learning Hierarchical Semantic Image Manipulation through Structured Representations

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #62

Seunghoon Hong · Xinchen Yan · Honglak Lee · Thomas Huang

Understanding, reasoning, and manipulating semantic concepts of images have been a fundamental research problem for decades. Previous work mainly focused on direct manipulation of natural image manifold through color strokes, key-points, textures, and holes-to-fill. In this work, we present a novel hierarchical framework for semantic image manipulation. Key to our hierarchical framework is that we employ structured semantic layout as our intermediate representations for manipulation. Initialized with coarse-level bounding boxes, our layout generator first creates pixel-wise semantic layout capturing the object shape, object-object interactions, and object-scene relations. Then our image generator fills in the pixel-level textures guided by the semantic layout. Such framework allows a user to manipulate images at object-level by adding, removing, and moving one bounding box at a time. Experimental evaluations demonstrate the advantages of the hierarchical manipulation framework over existing image generation and context hole-filing models, both qualitatively and quantitatively. Benefits of the hierarchical framework are further demonstrated in applications such as semantic object manipulation, interactive image editing, and data-driven image manipulation.

Cooperative Holistic Scene Understanding: Unifying 3D Object, Layout, and Camera Pose Estimation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #63

Siyuan Huang · Siyuan Qi · Yinxue Xiao · Yixin Zhu · Ying Nian Wu · Song-Chun Zhu

Holistic 3D indoor scene understanding refers to jointly recovering the i) object bounding boxes, ii) room layout, and iii) camera pose, all in 3D. The existing methods either are ineffective or only tackle the problem partially. In this paper, we propose an end-to-end model that simultaneously solves all three tasks in real-time given only a single RGB image. The essence of the proposed method is to improve the prediction by i) parametrizing the targets (e.g., 3D boxes) instead of directly estimating the targets, and ii) cooperative training across different modules in contrast to training these modules individually. Specifically, we parametrize the 3D object bounding boxes by the predictions from several modules, i.e., 3D camera pose and object attributes. The proposed method provides two major advantages: i) The parametrization helps maintain the consistency between the 2D image and the 3D world, thus largely reducing the prediction variances in 3D coordinates. ii) Constraints can be imposed on the parametrization to train different modules simultaneously. We call these constraints "cooperative losses" as they enable the joint training and inference. We employ three cooperative losses for 3D bounding boxes, 2D projections, and physical constraints to estimate a geometrically consistent and physically plausible 3D scene. Experiments on the SUN RGB-D dataset shows that the proposed method significantly outperforms prior approaches on 3D layout estimation, 3D object detection, 3D camera pose estimation, and holistic scene understanding.

3D-Aware Scene Manipulation via Inverse Graphics

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #64

Shunyu Yao · Tzu Ming Hsu · Jun-Yan Zhu · Jiajun Wu · Antonio Torralba · Bill Freeman · Josh Tenenbaum

We aim to obtain an interpretable, expressive, and disentangled scene representation that contains comprehensive structural and textural information for each object. Previous scene representations learned by neural networks are often uninterpretable, limited to a single object, or lacking 3D knowledge. In this work, we propose 3D scene de-rendering networks (3D-SDN) to address the above issues by integrating disentangled representations for semantics, geometry, and appearance into a deep generative model. Our scene encoder performs inverse graphics, translating a scene into a structured object-wise representation. Our decoder has two components: a differentiable shape renderer and a neural texture generator. The disentanglement of semantics, geometry, and appearance supports 3D-aware scene manipulation, e.g., rotating and moving objects freely while keeping the consistent shape and texture, and changing the object appearance without affecting its shape. Experiments demonstrate that our editing scheme based on 3D-SDN is superior to its 2D counterpart.

FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #65

Yixiao Ge · Zhuowan Li · Haiyu Zhao · Guojun Yin · Shuai Yi · Xiaogang Wang · hongsheng Li

Person re-identification (reID) is an important task that requires to retrieve a person's images from an image dataset, given one image of the person of interest. For learning robust person features, the pose variation of person images is one of the key challenges. Existing works targeting the problem either perform human alignment, or learn human-region-based representations. Extra pose information and computational cost is generally required for inference. To solve this issue, a Feature

Distilling Generative Adversarial Network (FD-GAN) is proposed for learning identity-related and pose-unrelated representations. It is a novel framework based on a Siamese structure with multiple novel discriminators on human poses and identities. In addition to the discriminators, a novel same-pose loss is also integrated, which requires appearance of a same person's generated images to be similar. After learning pose-unrelated person features with pose guidance, no auxiliary pose information and additional computational cost is required during testing. Our proposed FD-GAN achieves state-of-the-art performance on three person reID datasets, which demonstrates that the effectiveness and robust feature distilling capability of the proposed FD-GAN.

Sequence-to-Segment Networks for Segment Detection

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #66

Zijun Wei · Boyu Wang · Minh Hoai Nguyen · Jianming Zhang · Zhe Lin · Xiaohui Shen · Radomir Mech · Dimitris Samaras

Detecting segments of interest from an input sequence is a challenging problem which often requires not only good knowledge of individual target segments, but also contextual understanding of the entire input sequence and the relationships between the target segments. To address this problem, we propose the Sequence-to-Segment Network (S^{2}N), a novel end-to-end sequential encoder-decoder architecture. S^{2}N first encodes the input into a sequence of hidden states that progressively capture both local and holistic information. It then employs a novel decoding architecture, called Segment Detection Unit (SDU), that integrates the decoder state and encoder hidden states to detect segments sequentially. During training, we formulate the assignment of predicted segments to ground truth as bipartite matching and use the Earth Mover's Distance to calculate the localization errors. We experiment with S^{2}N on temporal action proposal generation and video summarization and show that S^{2}N achieves state-of-the-art performance on both tasks.

Stacked Semantics-Guided Attention Model for Fine-Grained Zero-Shot Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #67

yunlong yu · Zhong Ji · Yanwei Fu · Jichang Guo · Yanwei Pang · Zhongfei (Mark) Zhang

Zero-Shot Learning (ZSL) is generally achieved via aligning the semantic relationships between the visual features and the corresponding class semantic descriptions. However, using the global features to represent fine-grained images may lead to sub-optimal results since they neglect the discriminative differences of local regions. Besides, different regions contain distinct discriminative information. The important regions should contribute more to the prediction. To this end, we propose a novel stacked semantics-guided attention (S2GA) model to obtain semantic relevant features by using individual class semantic features to progressively guide the visual features to generate an attention map for weighting the importance of different local regions. Feeding both the integrated visual features and the class semantic features into a multi-class classification architecture, the proposed framework can be trained end-to-end. Extensive experimental results on CUB and NABird datasets show that the proposed approach has a consistent improvement on both fine-grained zero-shot classification and retrieval tasks.

DeepExposure: Learning to Expose Photos with Asynchronously Reinforced Adversarial Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #68

Runsheng Yu · Wenyu Liu · Yasen Zhang · Zhi Qu · Deli Zhao · Bo Zhang

The accurate exposure is the key of capturing high-quality photos in computational photography, especially for mobile phones that are limited by sizes of camera modules. Inspired by luminosity masks usually applied by professional photographers, in this paper, we develop a novel algorithm for learning local exposures with deep reinforcement adversarial learning. To be specific, we segment an image into sub-images that can reflect variations of dynamic range exposures according to raw low-level features. Based on these sub-images, a local exposure for each sub-image is automatically learned by virtue of policy network sequentially while the reward of learning is globally designed for striking a balance of overall exposures. The aesthetic evaluation function is approximated by discriminator in generative adversarial networks. The reinforcement learning and the adversarial learning are trained collaboratively by asynchronous deterministic policy gradient and generative loss approximation. To further simply the algorithmic architecture, we also prove the feasibility of leveraging the discriminator as the value function. Further more, we employ each local exposure to retouch the raw input image respectively, thus delivering multiple retouched images under different exposures which are fused with exposure blending. The extensive experiments verify that our algorithms are superior to state-of-the-art methods in terms of quantitative accuracy and visual illustration.

Self-Erasing Network for Integral Object Attention

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #69

Qibin Hou · PengTao Jiang · Yunchao Wei · Ming-Ming Cheng

Recently, adversarial erasing for weakly-supervised object attention has been deeply studied due to its capability in localizing integral object regions. However, such a strategy raises one key problem that attention regions will gradually expand to non-object regions as training iterations continue, which significantly decreases the quality of the produced attention maps. To tackle such an issue as well as promote the quality of object attention, we introduce a simple yet effective Self-Erasing Network (SeeNet) to prohibit attentions from spreading to unexpected background regions. In particular, SeeNet leverages two self-erasing strategies to encourage networks to use reliable object and background cues for learning to attention. In this way, integral object regions can be effectively highlighted without including much more background regions. To test the quality of the generated attention maps, we employ the mined object regions as heuristic cues for learning semantic segmentation models. Experiments on Pascal VOC well demonstrate the superiority of our SeeNet over other state-of-the-art methods.

Searching for Efficient Multi-Scale Architectures for Dense Image Prediction

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #70

Liang-Chieh Chen · Maxwell Collins · Yukun Zhu · George Papandreou · Barret Zoph · Florian Schroff · Hartwig Adam · Jon Shlens

The design of neural network architectures is an important component for achieving state-of-the-art performance with machine learning systems across a broad array of tasks. Much work has endeavored to design and build architectures automatically through clever construction of a search space paired with simple learning algorithms. Recent progress has demonstrated that such meta-learning methods may exceed scalable human-invented architectures on image classification tasks. An open question is the degree to which such methods may generalize to new domains. In this work we explore the construction of meta-learning techniques for dense image prediction focused on the tasks of scene parsing, person-part segmentation, and semantic image segmentation. Constructing viable search spaces in this domain is challenging because of the multi-scale representation of visual information and the necessity to operate on high resolution imagery. Based on a survey of techniques in dense image prediction, we construct a recursive search space and demonstrate that even with efficient random search, we can identify architectures that outperform human-invented architectures and achieve state-of-the-art performance on three dense prediction tasks including 82.7% on Cityscapes (street scene parsing), 71.3% on PASCAL-Person-Part (person-part segmentation), and 87.9% on PASCAL VOC 2012 (semantic image segmentation). Additionally, the resulting architecture is more computationally efficient, requiring half the parameters and half the computational cost as previous state of the art systems.

DifNet: Semantic Segmentation by Diffusion Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #71

Peng Jiang · Fanglin Gu · Yunhai Wang · Changhe Tu · Baoquan Chen

Deep Neural Networks (DNNs) have recently shown state of the art performance on semantic segmentation tasks, however, they still suffer from problems of poor boundary localization and spatial fragmented predictions. The difficulties lie in the requirement of making dense predictions from a long path model all at once since details are hard to keep when data goes through deeper layers. Instead, in this work, we decompose this difficult task into two relative simple sub-tasks: seed detection which is required to predict initial predictions without the need of wholeness and preciseness, and similarity estimation which measures the possibility of any two nodes belong to the same class without the need of knowing which class they are. We use one branch network for one sub-task each, and apply a cascade of random walks base on hierarchical semantics to approximate a complex diffusion process which propagates seed information to the whole image according to the estimated similarities. The proposed DifNet consistently produces improvements over the baseline models with the same depth and with the equivalent number of parameters, and also achieves promising performance on Pascal VOC and Pascal Context dataset. OurDifNet is trained end-to-end without complex loss functions.

Learning a High Fidelity Pose Invariant Model for High-resolution Face Frontalization

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #72

Jie Cao · Yibo Hu · Hongwen Zhang · Ran He · Zhenan Sun

Face frontalization refers to the process of synthesizing the frontal view of a face from a given profile. Due to self-occlusion and appearance distortion in the wild, it is extremely challenging to recover faithful results and preserve texture details in a high-resolution. This paper proposes a High Fidelity Pose Invariant Model (HF-PIM) to produce photographic and identity-preserving results. HF-PIM frontalizes the profiles through a novel texture warping procedure and leverages a dense correspondence field to bind the 2D and 3D surface spaces. We decompose the prerequisite of warping into dense correspondence field estimation and facial texture map recovering, which are both well addressed by deep networks. Different from those reconstruction methods relying on 3D data, we also propose Adversarial Residual Dictionary Learning (ARDL) to supervise facial texture map recovering with only monocular images. Exhaustive experiments on both controlled and uncontrolled environments demonstrate that the proposed method not only boosts the performance of pose-invariant face recognition but also dramatically improves high-resolution frontalization appearances.

Attention in Convolutional LSTM for Gesture Recognition

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #73

Liang Zhang · Guangming Zhu · Lin Mei · Peiyi Shen · Syed Afaq Ali Shah · Mohammed Bennamoun

Convolutional long short-term memory (LSTM) networks have been widely used for action/gesture recognition, and different attention mechanisms have also been embedded into the LSTM or the convolutional LSTM (ConvLSTM) networks. Based on the previous gesture recognition architectures which combine the three-dimensional convolution neural network (3DCNN) and ConvLSTM, this paper explores the effects of attention mechanism in ConvLSTM. Several variants of ConvLSTM are evaluated: (a) Removing the convolutional structures of the three gates in ConvLSTM, (b) Applying the attention mechanism on the input of ConvLSTM, (c) Reconstructing the input and (d) output gates respectively with the modified channel-wise attention mechanism. The evaluation results demonstrate that the spatial convolutions in the three gates scarcely contribute to the spatiotemporal feature fusion, and the attention mechanisms embedded into the input and output gates cannot improve the feature fusion. In other words, ConvLSTM mainly contributes to the temporal fusion along with the recurrent steps to learn the long-term spatiotemporal features, when taking as input the spatial or spatiotemporal features. On this basis, a new variant of LSTM is derived, in which the convolutional structures are only embedded into the input-to-state transition of LSTM. The code of the LSTM variants is publicly available.

Partially-Supervised Image Captioning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #74

Peter Anderson · Stephen Gould · Mark Johnson

Image captioning models are becoming increasingly successful at describing the content of images in restricted domains. However, if these models are to function in the wild --- for example, as assistants for people with impaired vision --- a much larger number and variety of visual concepts must be understood. To address this problem, we teach image captioning models new visual concepts from labeled images and object detection datasets. Since image labels and object classes can be interpreted as partial captions, we formulate this problem as learning from partially-specified sequence data. We then propose a novel algorithm for training sequence models, such as recurrent neural networks, on partially-specified sequences which we represent using finite state automata. In the context of image captioning, our method lifts the restriction that previously required image captioning models to be trained on paired image-sentence corpora only, or otherwise required specialized model architectures to take advantage of alternative data modalities. Applying our approach to an existing neural captioning model, we achieve state of the art results on the novel object captioning task using the COCO dataset. We further show that we can train a captioning model to describe new visual concepts from the Open Images dataset while maintaining competitive COCO evaluation scores.

Learning to Specialize with Knowledge Distillation for Visual Question Answering

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #75

Jonghwan Mun · Kimin Lee · Jinwoo Shin · Bohyung Han

Visual Question Answering (VQA) is a notoriously challenging problem because it involves various heterogeneous tasks defined by questions within a unified framework. Learning specialized models for individual types of tasks is intuitively attracting but surprisingly difficult; it is not straightforward to outperform naive independent ensemble approach. We present a principled algorithm to learn specialized models with knowledge distillation under a multiple choice learning (MCL) framework, where training examples are assigned dynamically to a subset of models for updating network parameters. The assigned and non-assigned models are learned to predict ground-truth answers and imitate their own base models before specialization, respectively. Our approach alleviates the limitation of data deficiency in existing MCL frameworks, and allows each model to learn its own specialized expertise without forgetting general knowledge. The proposed framework is model-agnostic and applicable to any tasks other than VQA, e.g., image classification with a large number of labels but few per-class examples, which is known to be difficult under existing MCL schemes. Our experimental results indeed demonstrate that our method outperforms other baselines for VQA and image classification.

Chain of Reasoning for Visual Question Answering

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #76

Chenfei Wu · Jinlai Liu · Xiaojie Wang · Xuan Dong

Reasoning plays an essential role in Visual Question Answering (VQA). Multi-step and dynamic reasoning is often necessary for answering complex questions. For example, a question "What is placed next to the bus on the right of the picture?" talks about a compound object "bus on the right," which is generated by the relation . Furthermore, a new relation including this compound object is then required to infer the answer. However, previous methods support either one-step or static reasoning, without updating relations or generating compound objects. This paper proposes a novel reasoning model for addressing these problems. A chain of reasoning (CoR) is constructed for supporting multi-step and dynamic reasoning on changed relations and objects. In detail, iteratively, the relational reasoning operations form new relations between objects, and the object refining operations generate new compound objects from relations. We achieve new state-of-the-art results on four publicly available datasets. The visualization of the chain of reasoning illustrates the progress that the CoR generates new compound objects that lead to the answer of the question step by step.

Learning Conditioned Graph Structures for Interpretable Visual Question Answering

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #77

Will Norcliffe-Brown · Stathis Vafeias · Sarah Parisot

Visual Question answering is a challenging problem requiring a combination of concepts from Computer Vision and Natural Language Processing. Most existing approaches use a two streams strategy, computing image and question features that are consequently merged using a variety of techniques. Nonetheless, very few rely on higher level image representations, which can capture semantic and spatial relationships. In this paper, we propose a novel graph-based approach for Visual Question Answering. Our method combines a graph learner module, which learns a question specific graph representation of the input image, with the recent concept of graph convolutions, aiming to learn image representations that capture question specific interactions. We test our approach on the VQA v2 dataset using a simple baseline architecture enhanced by the proposed graph learner module. We obtain promising results with 66.18% accuracy and demonstrate the interpretability of the proposed method. Code can be found at github.com/aimbrain/vqa-project.

Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #78

Medhini Narasimhan · Svetlana Lazebnik · Alexander Schwing

Accurately answering a question about a given image requires combining observations with general knowledge. While this is effortless for humans, reasoning with general knowledge remains an algorithmic challenge. To advance research in this direction a novel fact-based' visual question answering (FVQA) task has been introduced recently along with a large set of curated facts which link two entities, i.e., two possible answers, via a relation. Given a question-image pair, deep network techniques have been employed to successively reduce the large set of facts until one of the two entities of the final remaining fact is predicted as the answer. We observe that a successive process which considers one fact at a time to form a local decision is sub-optimal. Instead, we develop an entity graph and use a graph convolutional network toreason' about the correct answer by jointly considering all entities. We show on the challenging FVQA dataset that this leads to an improvement in accuracy of around 7% compared to the state-of-the-art.

Overcoming Language Priors in Visual Question Answering with Adversarial Regularization

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #79

Sainandan Ramakrishnan · Aishwarya Agrawal · Stefan Lee

Modern Visual Question Answering (VQA) models have been shown to rely heavily on superficial correlations between question and answer words learned during training -- \eg overwhelmingly reporting the type of room as kitchen or the sport being played as tennis, irrespective of the image. Most alarmingly, this shortcoming is often not well reflected during evaluation because the same strong priors exist in test distributions; however, a VQA system that fails to ground questions in image content would likely perform poorly in real-world settings. In this work, we present a novel regularization scheme for VQA that reduces this effect. We introduce a question-only model that takes as input the question encoding from the VQA model and must leverage language biases in order to succeed. We then pose training as an adversarial game between the VQA model and this question-only adversary -- discouraging the VQA model from capturing language biases in its question encoding.Further, we leverage this question-only model to estimate the mutual information between the image and answer given the question, which we maximize explicitly to encourage visual grounding. Our approach is a model agnostic training procedure and simple to implement. We show empirically that it can improve performance significantly on a bias-sensitive split of the VQA dataset for multiple base models -- achieving state-of-the-art on this task. Further, on standard VQA tasks, our approach shows significantly less drop in accuracy compared to existing bias-reducing VQA models.

Non-Local Recurrent Network for Image Restoration

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #80

Ding Liu · Bihan Wen · Yuchen Fan · Chen Change Loy · Thomas Huang

Many classic methods have shown non-local self-similarity in natural images to be an effective prior for image restoration. However, it remains unclear and challenging to make use of this intrinsic property via deep networks. In this paper, we propose a non-local recurrent network (NLRN) as the first attempt to incorporate non-local operations into a recurrent neural network (RNN) for image restoration. The main contributions of this work are: (1) Unlike existing methods that measure self-similarity in an isolated manner, the proposed non-local module can be flexibly integrated into existing deep networks for end-to-end training to capture deep feature correlation between each location and its neighborhood. (2) We fully employ the RNN structure for its parameter efficiency and allow deep feature correlation to be propagated along adjacent recurrent states. This new design boosts robustness against inaccurate correlation estimation due to severely degraded images. (3) We show that it is essential to maintain a confined neighborhood for computing deep feature correlation given degraded images. This is in contrast to existing practice that deploys the whole image. Extensive experiments on both image denoising and super-resolution tasks are conducted. Thanks to the recurrent non-local operations and correlation propagation, the proposed NLRN achieves superior results to state-of-the-art methods with many fewer parameters.

Neural Nearest Neighbors Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #81

Tobias Plötz · Stefan Roth

Non-local methods exploiting the self-similarity of natural signals have been well studied, for example in image analysis and restoration. Existing approaches, however, rely on k-nearest neighbors (KNN) matching in a fixed feature space. The main hurdle in optimizing this feature space w.r.t. application performance is the non-differentiability of the KNN selection rule. To overcome this, we propose a continuous deterministic relaxation of KNN selection that maintains differentiability w.r.t. pairwise distances, but retains the original KNN as the limit of a temperature parameter approaching zero. To exploit our relaxation, we propose the neural nearest neighbors block (N3 block), a novel non-local processing layer that leverages the principle of self-similarity and can be used as building block in modern neural network architectures. We show its effectiveness for the set reasoning task of correspondence classification as well as for image restoration, including image denoising and single image super-resolution, where we outperform strong convolutional neural network (CNN) baselines and recent non-local models that rely on KNN selection in hand-chosen features spaces.

Training deep learning based denoisers without ground truth data

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #82

Shakarim Soltanayev · Se Young Chun

Recently developed deep-learning-based denoisers often outperform state-of-the-art conventional denoisers, such as the BM3D. They are typically trained to minimizethe mean squared error (MSE) between the output image of a deep neural networkand a ground truth image. In deep learning based denoisers, it is important to use high quality noiseless ground truth data for high performance, but it is often challenging or even infeasible to obtain noiseless images in application areas such as hyperspectral remote sensing and medical imaging. In this article, we propose a method based on Stein’s unbiased risk estimator (SURE) for training deep neural network denoisers only based on the use of noisy images. We demonstrate that our SURE-based method, without the use of ground truth data, is able to train deep neural network denoisers to yield performances close to those networks trained with ground truth, and to outperform the state-of-the-art denoiser BM3D. Further improvements were achieved when noisy test images were used for training of denoiser networks using our proposed SURE-based method.

Adversarial Regularizers in Inverse Problems

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #83

Sebastian Lunz · Carola Schoenlieb · Ozan Öktem

Inverse Problems in medical imaging and computer vision are traditionally solved using purely model-based methods. Among those variational regularization models are one of the most popular approaches. We propose a new framework for applying data-driven approaches to inverse problems, using a neural network as a regularization functional. The network learns to discriminate between the distribution of ground truth images and the distribution of unregularized reconstructions. Once trained, the network is applied to the inverse problem by solving the corresponding variational problem. Unlike other data-based approaches for inverse problems, the algorithm can be applied even if only unsupervised training data is available. Experiments demonstrate the potential of the framework for denoising on the BSDS dataset and for computer tomography reconstruction on the LIDC dataset.

Densely Connected Attention Propagation for Reading Comprehension

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #84

Yi Tay · Anh Tuan Luu · Siu Cheung Hui · Jian Su

We propose DecaProp (Densely Connected Attention Propagation), a new densely connected neural architecture for reading comprehension (RC). There are two distinct characteristics of our model. Firstly, our model densely connects all pairwise layers of the network, modeling relationships between passage and query across all hierarchical levels. Secondly, the dense connectors in our network are learned via attention instead of standard residual skip-connectors. To this end, we propose novel Bidirectional Attention Connectors (BAC) for efficiently forging connections throughout the network. We conduct extensive experiments on four challenging RC benchmarks. Our proposed approach achieves state-of-the-art results on all four, outperforming existing baselines by up to 2.6% to 14.2% in absolute F1 score.

Layer-Wise Coordination between Encoder and Decoder for Neural Machine Translation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #85

Tianyu He · Xu Tan · Yingce Xia · Di He · Tao Qin · Zhibo Chen · Tie-Yan Liu

Neural Machine Translation (NMT) has achieved remarkable progress with the quick evolvement of model structures. In this paper, we propose the concept of layer-wise coordination for NMT, which explicitly coordinates the learning of hidden representations of the encoder and decoder together layer by layer, gradually from low level to high level. Specifically, we design a layer-wise attention and mixed attention mechanism, and further share the parameters of each layer between the encoder and decoder to regularize and coordinate the learning. Experiments show that combined with the state-of-the-art Transformer model, layer-wise coordination achieves improvements on three IWSLT and two WMT translation tasks. More specifically, our method achieves 34.43 and 29.01 BLEU score on WMT16 English-Romanian and WMT14 English-German tasks, outperforming the Transformer baseline.

---

e-SNLI: Natural Language Inference with Natural Language Explanations

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #86

Oana-Maria Camburu · Tim Rocktäschel · Thomas Lukasiewicz · Phil Blunsom

In order for machine learning to garner widespread public adoption, models must be able to provide interpretable and robust explanations for their decisions, as well as learn from human-provided explanations at train time. In this work, we extend the Stanford Natural Language Inference dataset with an additional layer of human-annotated natural language explanations of the entailment relations. We further implement models that incorporate these explanations into their training process and output them at test time. We show how our corpus of explanations, which we call e-SNLI, can be used for various goals, such as obtaining full sentence justifications of a model’s decisions, improving universal sentence representations and transferring to out-of-domain NLI datasets. Our dataset thus opens up a range of research directions for using natural language explanations, both for improving models and for asserting their trust

---

The challenge of realistic music generation: modelling raw audio at scale

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #87

Sander Dieleman · Aaron van den Oord · Karen Simonyan

Realistic music generation is a challenging task. When building generative models of music that are learnt from data, typically high-level representations such as scores or MIDI are used that abstract away the idiosyncrasies of a particular performance. But these nuances are very important for our perception of musicality and realism, so in this work we embark on modelling music in the raw audio domain. It has been shown that autoregressive models excel at generating raw audio waveforms of speech, but when applied to music, we find them biased towards capturing local signal structure at the expense of modelling long-range correlations. This is problematic because music exhibits structure at many different timescales. In this work, we explore autoregressive discrete autoencoders (ADAs) as a means to enable autoregressive models to capture long-range correlations in waveforms. We find that they allow us to unconditionally generate piano music directly in the raw audio domain, which shows stylistic consistency across tens of seconds.

Fully Neural Network Based Speech Recognition on Mobile and Embedded Devices

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #88

Jinhwan Park · Yoonho Boo · Iksoo Choi · Sungho Shin · Wonyong Sung

Real-time automatic speech recognition (ASR) on mobile and embedded devices has been of great interests for many years. We present real-time speech recognition on smartphones or embedded systems by employing recurrent neural network (RNN) based acoustic models, RNN based language models, and beam-search decoding. The acoustic model is end-to-end trained with connectionist temporal classification (CTC) loss. The RNN implementation on embedded devices can suffer from excessive DRAM accesses because the parameter size of a neural network usually exceeds that of the cache memory and the parameters are used only once for each time step. To remedy this problem, we employ a multi-time step parallelization approach that computes multiple output samples at a time with the parameters fetched from the DRAM. Since the number of DRAM accesses can be reduced in proportion to the number of parallelization steps, we can achieve a high processing speed. However, conventional RNNs, such as long short-term memory (LSTM) or gated recurrent unit (GRU), do not permit multi-time step parallelization. We construct an acoustic model by combining simple recurrent units (SRUs) and depth-wise 1-dimensional convolution layers for multi-time step parallelization. Both the character and word piece models are developed for acoustic modeling, and the corresponding RNN based language models are used for beam search decoding. We achieve a competitive WER for WSJ corpus using the entire model size of around 15MB and achieve real-time speed using only a single core ARM without GPU or special hardware.

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #89

Ye Jia · Yu Zhang · Ron Weiss · Quan Wang · Jonathan Shen · Fei Ren · zhifeng Chen · Patrick Nguyen · Ruoming Pang · Ignacio Lopez Moreno · Yonghui Wu

We describe a neural network-based system for text-to-speech (TTS) synthesis that is able to generate speech audio in the voice of many different speakers, including those unseen during training. Our system consists of three independently trained components: (1) a speaker encoder network, trained on a speaker verification task using an independent dataset of noisy speech from thousands of speakers without transcripts, to generate a fixed-dimensional embedding vector from seconds of reference speech from a target speaker; (2) a sequence-to-sequence synthesis network based on Tacotron 2, which generates a mel spectrogram from text, conditioned on the speaker embedding; (3) an auto-regressive WaveNet-based vocoder that converts the mel spectrogram into a sequence of time domain waveform samples. We demonstrate that the proposed model is able to transfer the knowledge of speaker variability learned by the discriminatively-trained speaker encoder to the new task, and is able to synthesize natural speech from speakers that were not seen during training. We quantify the importance of training the speaker encoder on a large and diverse speaker set in order to obtain the best generalization performance. Finally, we show that randomly sampled speaker embeddings can be used to synthesize speech in the voice of novel speakers dissimilar from those used in training, indicating that the model has learned a high quality speaker representation.

SING: Symbol-to-Instrument Neural Generator

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #90

Alexandre Defossez · Neil Zeghidour · Nicolas Usunier · Leon Bottou · Francis Bach

Recent progress in deep learning for audio synthesis opens the way to models that directly produce the waveform, shifting away from the traditional paradigm of relying on vocoders or MIDI synthesizers for speech or music generation. Despite their successes, current state-of-the-art neural audio synthesizers such as WaveNet and SampleRNN suffer from prohibitive training and inference times because they are based on autoregressive models that generate audio samples one at a time at a rate of 16kHz. In this work, we study the more computationally efficient alternative of generating the waveform frame-by-frame with large strides. We present a lightweight neural audio synthesizer for the original task of generating musical notes given desired instrument, pitch and velocity. Our model is trained end-to-end to generate notes from nearly 1000 instruments with a single decoder, thanks to a new loss function that minimizes the distances between the log spectrograms of the generated and target waveforms. On the generalization task of synthesizing notes for pairs of pitch

and instrument not seen during training, SING produces audio with significantly improved perceptual quality compared to a state-of-the-art autoencoder based on WaveNet as measured by a Mean Opinion Score (MOS), and is about 32 times faster for training and 2, 500 times faster for inference.

Neural Voice Cloning with a Few Samples

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #91

Sercan Arik · Jitong Chen · Kainan Peng · Wei Ping · Yanqi Zhou

Voice cloning is a highly desired feature for personalized speech interfaces. We introduce a neural voice cloning system that learns to synthesize a person's voice from only a few audio samples. We study two approaches: speaker adaptation and speaker encoding. Speaker adaptation is based on fine-tuning a multi-speaker generative model. Speaker encoding is based on training a separate model to directly infer a new speaker embedding, which will be applied to a multi-speaker generative model. In terms of naturalness of the speech and similarity to the original speaker, both approaches can achieve good performance, even with a few cloning audios. While speaker adaptation can achieve slightly better naturalness and similarity, cloning time and required memory for the speaker encoding approach are significantly less, making it more favorable for low-resource deployment.

GroupReduce: Block-Wise Low-Rank Approximation for Neural Language Model Shrinking

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #92

Patrick Chen · Si Si · Yang Li · Ciprian Chelba · Cho-Jui Hsieh

Model compression is essential for serving large deep neural nets on devices with limited resources or applications that require real-time responses. For advanced NLP problems, a neural language model usually consists of recurrent layers (e.g., using LSTM cells), an embedding matrix for representing input tokens, and a softmax layer for generating output tokens. For problems with a very large vocabulary size, the embedding and the softmax matrices can account for more than half of the model size. For instance, the bigLSTM model achieves state-of-the-art performance on the One-Billion-Word (OBW) dataset with around 800k vocabulary, and its word embedding and softmax matrices use more than 6GBytes space, and are responsible for over 90\% of the model parameters. In this paper, we propose GroupReduce, a novel compression method for neural language models, based on vocabulary-partition (block) based low-rank matrix approximation and the inherent frequency distribution of tokens (the power-law distribution of words). We start by grouping words into c blocks based on their frequency, and then refine the clustering iteratively by constructing weighted low-rank approximation for each block, where the weights are based the frequencies of the words in the block. The experimental results show our method can significantly outperform traditional compression methods such as low-rank approximation and pruning. On the OBW dataset, our method achieved 6.6x compression rate for the embedding and softmax matrices, and when combined with quantization, our method can achieve 26x compression rate without losing prediction accuracy.

Dialog-to-Action: Conversational Question Answering Over a Large-Scale Knowledge Base

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #93

Daya Guo · Duyu Tang · Nan Duan · Ming Zhou · Jian Yin

We present an approach to map utterances in conversation to logical forms, which will be executed on a large-scale knowledge base. To handle enormous ellipsis phenomena in conversation, we introduce dialog memory management to manipulate historical entities, predicates, and logical forms when inferring the logical form of current utterances. Dialog memory management is embodied in a generative model, in which a logical form is interpreted in a top-down manner following a small and flexible grammar. We learn the model from denotations without explicit annotation of logical forms, and evaluate it on a large-scale dataset consisting of 200K dialogs over 12.8M entities. Results verify the benefits of modeling dialog memory, and show that our semantic parsing-based approach outperforms a memory network based encoder-decoder model by a huge margin.

Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #94

Yizhe Zhang · Michel Galley · Jianfeng Gao · Zhe Gan · Xiujun Li · Chris Brockett · Bill Dolan

Responses generated by neural conversational models tend to lack informativeness and diversity. We present Adversarial Information Maximization (AIM), an adversarial learning framework that addresses these two related but distinct problems. To foster response diversity, we leverage adversarial training that allows distributional matching of synthetic and real responses. To improve informativeness, our framework explicitly optimizes a variational lower bound on pairwise mutual information between query and response. Empirical results from automatic and human evaluations demonstrate that our methods significantly boost informativeness and diversity.

Answerer in Questioner's Mind: Information Theoretic Approach to Goal-Oriented Visual Dialog

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #95

Sang-Woo Lee · Yu-Jung Heo · Byoung-Tak Zhang

Goal-oriented dialog has been given attention due to its numerous applications in artificial intelligence. Goal-oriented dialogue tasks occur when a questioner asks an action-oriented question and an answerer responds with the intent of letting the questioner know a correct action to take. To ask the adequate question, deep learning and reinforcement learning have been recently applied. However, these approaches struggle to find a competent recurrent neural questioner, owing to the complexity of learning a series of sentences. Motivated by theory of mind, we propose "Answerer in Questioner's Mind" (AQM), a novel information theoretic algorithm for goal-oriented dialog. With AQM, a questioner asks and infers based on an approximated probabilistic model of the answerer. The questioner figures out the answerer’s intention via selecting a plausible question by explicitly calculating the information gain of the candidate intentions and possible answers to each question. We test our framework on two goal-oriented visual dialog tasks: "MNIST Counting Dialog" and "GuessWhat?!". In our experiments, AQM outperforms comparative algorithms by a large margin.

Trajectory Convolution for Action Recognition

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #96

Yue Zhao · Yuanjun Xiong · Dahua Lin

How to leverage the temporal dimension is a key question in video analysis. Recent works suggest an efficient approach to video feature learning, i.e., factorizing 3D convolutions into separate components respectively for spatial and temporal convolutions. The temporal convolution, however, comes with an implicit assumption – the feature maps across time steps are well aligned so that the features at the same locations can be aggregated. This assumption may be overly strong in practical applications, especially in action recognition where the motion serves as a crucial cue. In this work, we propose a new CNN architecture TrajectoryNet, which incorporates trajectory convolution, a new operation for integrating features along the temporal dimension, to replace the existing temporal convolution. This operation explicitly takes into account the changes in contents caused by deformation or motion, allowing the visual features to be aggregated along the the motion paths, trajectories. On two large-scale action recognition datasets, namely, Something-Something and Kinetics, the proposed network architecture achieves notable improvement over strong baselines.

Video Prediction via Selective Sampling

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #97

Jingwei Xu · Bingbing Ni · Xiaokang Yang

Most adversarial learning based video prediction methods suffer from image blur, since the commonly used adversarial and regression loss pair work rather in a competitive way than collaboration, yielding compromised blur effect. In the meantime, as often relying on a single-pass architecture, the predictor is inadequate to explicitly capture the forthcoming uncertainty. Our work involves two key insights: (1) Video prediction can be approached as a stochastic process: we sample a collection of proposals conforming to possible frame distribution at following time stamp, and one can select the final prediction from it. (2) De-coupling combined loss functions into dedicatedly designed sub-networks encourages them to work in a collaborative way. Combining above two insights we propose a two-stage network called VPSS (\textbf{V}ideo \textbf{P}rediction via \textbf{S}elective \textbf{S}ampling). Specifically a \emph{Sampling} module produces a collection of high quality proposals, facilitated by a multiple choice adversarial learning scheme, yielding diverse frame proposal set. Subsequently a \emph{Selection} module selects high possibility candidates from proposals and combines them to produce final prediction. Extensive experiments on diverse challenging datasets demonstrate the effectiveness of proposed video prediction approach, i.e., yielding more diverse proposals and accurate prediction results.

Unsupervised Learning of Artistic Styles with Archetypal Style Analysis

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #98

Daan Wynen · Cordelia Schmid · Julien Mairal

In this paper, we introduce an unsupervised learning approach to automatically dis- cover, summarize, and manipulate artistic styles from large collections of paintings. Our method is based on archetypal analysis, which is an unsupervised learning technique akin to sparse coding with a geometric interpretation. When applied to deep image representations from a data collection, it learns a dictionary of archetypal styles, which can be easily visualized. After training the model, the style of a new image, which is characterized by local statistics of deep visual features, is approximated by a sparse convex combination of archetypes. This allows us to interpret which archetypal styles are present in the input image, and in which proportion. Finally, our approach allows us to manipulate the coefficients of the latent archetypal decomposition, and achieve various special effects such as style enhancement, transfer, and interpolation between multiple archetypes.

Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #99

Guanhong Tao · Shiqing Ma · Yingqi Liu · Xiangyu Zhang

Adversarial sample attacks perturb benign inputs to induce DNN misbehaviors. Recent research has demonstrated the widespread presence and the devastating consequences of such attacks. Existing defense techniques either assume prior knowledge of specific attacks or may not work well on complex models due to their underlying assumptions. We argue that adversarial sample attacks are deeply entangled with interpretability of DNN models: while classification results on benign inputs can be reasoned based on the human perceptible features/attributes, results on adversarial samples can hardly be explained. Therefore, we propose a novel adversarial sample detection technique for face recognition models, based on interpretability. It features a novel bi-directional correspondence inference between attributes and internal neurons to identify neurons critical for individual attributes. The activation values of critical neurons are enhanced to amplify the reasoning part of the computation and the values of other neurons are weakened to suppress the uninterpretable part. The classification results after such transformation are compared with those of the original model to detect adversaries. Results show that our technique can achieve 94% detection accuracy for 7 different kinds of attacks with 9.91% false positives on benign inputs. In contrast, a state-of-the-art feature squeezing technique can only achieve 55% accuracy with 23.3% false positives.

Speaker-Follower Models for Vision-and-Language Navigation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #100

Daniel Fried · Ronghang Hu · Volkan Cirik · Anna Rohrbach · Jacob Andreas · Louis-Philippe Morency · Taylor Berg-Kirkpatrick · Kate Saenko · Dan Klein · Trevor Darrell

Navigation guided by natural language instructions presents a challenging reasoning problem for instruction followers. Natural language instructions typically identify only a few high-level decisions and landmarks rather than complete low-level motor behaviors; much of the missing information must be inferred based on perceptual context. In machine learning settings, this is doubly challenging: it is difficult to collect enough annotated data to enable learning of this reasoning process from scratch, and also difficult to implement the reasoning process using generic sequence models. Here we describe an approach to vision-and-language navigation that addresses both these issues with an embedded speaker model. We use this speaker model to (1) synthesize new instructions for data augmentation and to (2) implement pragmatic reasoning, which evaluates how well candidate action sequences explain an instruction. Both steps are supported by a panoramic action space that reflects the granularity of human-generated instructions. Experiments show that all three components of this approach---speaker-driven data augmentation, pragmatic reasoning and panoramic action space---dramatically improve the performance of a baseline instruction follower, more than doubling the success rate over the best existing approach on a standard benchmark.

Neural Code Comprehension: A Learnable Representation of Code Semantics

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #101

Tal Ben-Nun · Alice Shoshana Jakobovits · Torsten Hoefler

With the recent success of embeddings in natural language processing, research has been conducted into applying similar methods to code analysis. Most works attempt to process the code directly or use a syntactic tree representation, treating it like sentences written in a natural language. However, none of the existing methods are sufficient to comprehend program semantics robustly, due to structural features such as function calls, branching, and interchangeable order of statements. In this paper, we propose a novel processing technique to learn code semantics, and apply it to a variety of program analysis tasks. In particular, we stipulate that a robust distributional hypothesis of code applies to both human- and machine-generated programs. Following this hypothesis, we define an embedding space, inst2vec, based on an Intermediate Representation (IR) of the code that is independent of the source programming language. We provide a novel definition of contextual flow for this IR, leveraging both the underlying data- and control-flow of the program. We then analyze the embeddings qualitatively using analogies and clustering, and evaluate the learned representation on three different high-level tasks. We show that even without fine-tuning, a single RNN architecture and fixed inst2vec embeddings outperform specialized approaches for performance prediction (compute device mapping, optimal thread coarsening); and algorithm classification from raw code (104 classes), where we set a new state-of-the-art.

MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #102

Edward Choi · Cao Xiao · Walter Stewart · Jimeng Sun

Deep learning models exhibit state-of-the-art performance for many predictive healthcare tasks using electronic health records (EHR) data, but these models typically require training data volume that exceeds the capacity of most healthcare systems. External resources such as medical ontologies are used to bridge the data volume constraint, but this approach is often not directly applicable or useful because of inconsistencies with terminology. To solve the data insufficiency challenge, we leverage the inherent multilevel structure of EHR data and, in particular, the encoded relationships among medical codes. We propose Multilevel Medical Embedding (MiME) which learns the multilevel embedding of EHR data while jointly performing auxiliary prediction tasks that rely on this inherent EHR structure without the need for external labels. We conducted two prediction tasks, heart failure prediction and sequential disease prediction, where MiME outperformed baseline methods in diverse evaluation settings. In particular, MiME consistently outperformed all baselines when predicting heart failure on datasets of different volumes, especially demonstrating the greatest performance improvement (15% relative gain in PR-AUC over the best baseline) on the smallest dataset, demonstrating its ability to effectively model the multilevel structure of EHR data.

Distilled Wasserstein Learning for Word Embedding and Topic Modeling

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #103

Hongteng Xu · Wenlin Wang · Wei Liu · Lawrence Carin

We propose a novel Wasserstein method with a distillation mechanism, yielding joint learning of word embeddings and topics. The proposed method is based on the fact that the Euclidean distance between word embeddings may be employed as the underlying distance in the Wasserstein topic model. The word distributions of topics, their optimal transport to the word distributions of documents, and the embeddings of words are learned in a unified framework. When learning the topic model, we leverage a distilled ground-distance matrix to update the topic distributions and smoothly calculate the corresponding optimal transports. Such a strategy provides the updating of word embeddings with robust guidance, improving algorithm convergence. As an application, we focus on patient admission records, in which the proposed method embeds the codes of diseases and procedures and learns the topics of admissions, obtaining superior performance on clinically-meaningful disease network construction, mortality prediction as a function of admission codes, and procedure recommendation.

Learning to Optimize Tensor Programs

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #104

Tianqi Chen · Lianmin Zheng · Eddie Yan · Ziheng Jiang · Thierry Moreau · Luis Ceze · Carlos Guestrin · Arvind Krishnamurthy

We introduce a learning-based framework to optimize tensor programs for deep learning workloads. Efficient implementations of tensor operators, such as matrix multiplication and high dimensional convolution are key enablers of effective deep learning systems. However, existing systems rely on manually optimized libraries such as cuDNN where only a narrow range of server class GPUs are well-supported. The reliance on hardware specific operator libraries limits the applicability of high-level graph optimizations and incurs significant engineering costs when deploying to new hardware targets. We use learning to remove this engineering burden. We learn domain specific statistical cost models to guide the search of tensor operator implementations over billions of possible program variants. We further accelerate the search by effective model transfer across workloads. Experimental results show that our framework delivers performance competitive with state-of-the-art hand-tuned libraries for low-power CPU, mobile GPU, and server-class GPU.

Learning to Solve SMT Formulas

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #105

Mislav Balunovic · Pavol Bielik · Martin Vechev

We present a new approach for learning to solve SMT formulas. We phrase the challenge of solving SMT formulas as a tree search problem where at each step a transformation is applied to the input formula until the formula is solved. Our approach works in two phases: first, given a dataset of unsolved formulas we learn a policy that for each formula selects a suitable transformation to apply at each step in order to solve the formula, and second, we synthesize a strategy in the form of a loop-free program with branches. This strategy is an interpretable representation of the policy decisions and is used to guide the SMT solver to decide formulas more efficiently, without requiring any modification to the solver itself and without needing to evaluate the learned policy at inference time. We show that our approach is effective in practice - it solves 17% more formulas over a range of benchmarks and achieves up to 100x runtime improvement over a state-of-the-art SMT solver.

Data center cooling using model-predictive control

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #106

Nevena Lazic · Craig Boutilier · Tyler Lu · Eehern Wong · Binz Roy · MK Ryu · Greg Imwalle

Despite impressive recent advances in reinforcement learning (RL), its deployment in real-world physical systems is often complicated by unexpected events, limited data, and the potential for expensive failures. In this paper, we describe an application of RL “in the wild” to the task of regulating temperatures and airflow inside a large-scale data center (DC). Adopting a data-driven, model-based approach, we demonstrate that an RL agent with little prior knowledge is able to effectively and safely regulate conditions on a server floor after just a few hours of exploration, while improving operational efficiency relative to existing PID controllers.

Bayesian Inference of Temporal Task Specifications from Demonstrations

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #107

Ankit Shah · Pritish Kamath · Julie A Shah · Shen Li

When observing task demonstrations, human apprentices are able to identify whether a given task is executed correctly long before they gain expertise in actually performing that task. Prior research into learning from demonstrations (LfD) has failed to capture this notion of the acceptability of an execution; meanwhile, temporal logics provide a flexible language for expressing task specifications. Inspired by this, we present Bayesian specification inference, a probabilistic model for inferring task specification as a temporal logic formula. We incorporate methods from probabilistic programming to define our priors, along with a domain-independent likelihood function to enable sampling-based inference. We demonstrate the efficacy of our model for inferring true specifications with over 90% similarity between the inferred specification and the ground truth, both within a synthetic domain and a real-world table setting task.

Training Deep Neural Networks with 8-bit Floating Point Numbers

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #108

Naigang Wang · Jungwook Choi · Daniel Brand · Chia-Yu Chen · Kailash Gopalakrishnan

The state-of-the-art hardware platforms for training deep neural networks are moving from traditional single precision (32-bit) computations towards 16 bits of precision - in large part due to the high energy efficiency and smaller bit storage associated with using reduced-precision representations. However, unlike inference, training with numbers represented with less than 16 bits has been challenging due to the need to maintain fidelity of the gradient computations during back-propagation. Here we demonstrate, for the first time, the successful training of deep neural networks using 8-bit floating point numbers while fully maintaining the accuracy on a spectrum of deep learning models and datasets. In addition to reducing the data and computation precision to 8 bits, we also successfully reduce the arithmetic precision for additions (used in partial product accumulation and weight updates) from 32 bits to 16 bits through the introduction of a number of key ideas including chunk-based accumulation and floating point stochastic rounding. The use of these novel techniques lays the foundation for a new generation of hardware training platforms with the potential for 2-4 times improved throughput over today's systems.

Snap ML: A Hierarchical Framework for Machine Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #109

Celestine Dünner · Thomas Parnell · Dimitrios Sarigiannis · Nikolas Ioannou · Andreea Anghel · Gummadi Ravi · Madhusudanan Kandasamy · Haralampos Pozidis

We describe a new software framework for fast training of generalized linear models. The framework, named Snap Machine Learning (Snap ML), combines recent advances in machine learning systems and algorithms in a nested manner to reflect the hierarchical architecture of modern computing systems. We prove theoretically that such a hierarchical system can accelerate training in distributed environments where intra-node communication is cheaper than inter-node communication. Additionally, we provide a review of the implementation of Snap ML in terms of GPU acceleration, pipelining, communication patterns and software architecture, highlighting aspects that were critical for achieving high performance. We evaluate the performance of Snap ML in both single-node and multi-node environments, quantifying the benefit of the hierarchical scheme and the data streaming functionality, and comparing with other widely-used machine learning software frameworks. Finally, we present a logistic regression benchmark on the Criteo Terabyte Click Logs dataset and show that Snap ML achieves the same test loss an order of magnitude faster than any of the previously reported results, including those obtained using TensorFlow and scikit-learn.

Learning filter widths of spectral decompositions with wavelets

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #110

Haidar Khan · Bulent Yener

Time series classification using deep neural networks, such as convolutional neural networks (CNN), operate on the spectral decomposition of the time series computed using a preprocessing step. This step can include a large number of hyperparameters, such as window length, filter widths, and filter shapes, each with a range of possible values that must be chosen using time and data intensive cross-validation procedures. We propose the wavelet deconvolution (WD) layer as an efficient alternative to this preprocessing step that eliminates a significant number of hyperparameters. The WD layer uses wavelet functions with adjustable scale parameters to learn the spectral decomposition directly from the signal. Using backpropagation, we show the scale parameters can be optimized with gradient descent. Furthermore, the WD layer adds interpretability to the learned time series classifier by exploiting the properties of the wavelet transform. In our experiments, we show that the WD layer can automatically extract the frequency content used to generate a dataset. The WD layer combined with a CNN applied to the phone recognition task on the TIMIT database achieves a phone error rate of 18.1\%, a relative improvement of 4\% over the baseline CNN.

Experiments on a dataset where engineered features are not available showed WD+CNN is the best performing method. Our results show that the WD layer can improve neural network based time series classifiers both in accuracy and interpretability by learning directly from the input signal.

A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #111

Jeffrey Chan · Valerio Perrone · Jeffrey Spence · Paul Jenkins · Sara Mathieson · Yun Song

An explosion of high-throughput DNA sequencing in the past decade has led to a surge of interest in population-scale inference with whole-genome data. Recent work in population genetics has centered on designing inference methods for relatively simple model classes, and few scalable general-purpose inference techniques exist for more realistic, complex models. To achieve this, two inferential challenges need to be addressed: (1) population data are exchangeable, calling for methods that efficiently exploit the symmetries of the data, and (2) computing likelihoods is intractable as it requires integrating over a set of correlated, extremely high-dimensional latent variables. These challenges are traditionally tackled by likelihood-free methods that use scientific simulators to generate datasets and reduce them to hand-designed, permutation-invariant summary statistics, often leading to inaccurate inference. In this work, we develop an exchangeable neural network that performs summary statistic-free, likelihood-free inference. Our framework can be applied in a black-box fashion across a variety of simulation-based tasks, both within and outside biology. We demonstrate the power of our approach on the recombination hotspot testing problem, outperforming the state-of-the-art.

Latent Gaussian Activity Propagation: Using Smoothness and Structure to Separate and Localize Sounds in Large Noisy Environments

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #112

Daniel Johnson · Daniel Gorelik · Ross E Mawhorter · Kyle Suver · Weiqing Gu · Steven Xing · Cody Gabriel · Peter Sankhagowit

We present an approach for simultaneously separating and localizing multiple sound sources using recorded microphone data. Inspired by topic models, our approach is based on a probabilistic model of inter-microphone phase differences, and poses separation and localization as a Bayesian inference problem. We assume sound activity is locally smooth across time, frequency, and location, and use the known position of the microphones to obtain a consistent separation. We compare the performance of our method against existing algorithms on simulated anechoic voice data and find that it obtains high performance across a variety of input conditions.

Inferring Latent Velocities from Weather Radar Data using Gaussian Processes

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #113

Rico Angell · Daniel Sheldon

Archived data from the US network of weather radars hold detailed information about bird migration over the last 25 years, including very high-resolution partial measurements of velocity. Historically, most of this spatial resolution is discarded and velocities are summarized at a very small number of locations due to modeling and algorithmic limitations. This paper presents a Gaussian process (GP) model to reconstruct high-resolution full velocity fields across the entire US. The GP faithfully models all aspects of the problem in a single joint framework, including spatially random velocities, partial velocity measurements, station-specific geometries, measurement noise, and an ambiguity known as aliasing. We develop fast inference algorithms based on the FFT; to do so, we employ a creative use of Laplace's method to sidestep the fact that the kernel of the joint process is non-stationary.

Bayesian Nonparametric Spectral Estimation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #114

Felipe Tobar

Spectral estimation (SE) aims to identify how the energy of a signal (e.g., a time series) is distributed across different frequencies. This can become particularly challenging when only partial and noisy observations of the signal are available, where current methods fail to handle uncertainty appropriately. In this context, we propose a joint probabilistic model for signals, observations and spectra, where SE is addressed as an inference problem. Assuming a Gaussian process prior over the signal, we apply Bayes' rule to find the analytic posterior distribution of the spectrum given a set of observations. Besides its expressiveness and natural account of spectral uncertainty, the proposed model also provides a functional-form representation of the power spectral density, which can be optimised efficiently. Comparison with previous approaches is addressed theoretically, showing that the proposed method is an infinite-dimensional variant of the Lomb-Scargle approach, and also empirically through three experiments.

Learning a Warping Distance from Unlabeled Time Series Using Sequence Autoencoders

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #115

Abubakar Abid · James Zou

Measuring similarities between unlabeled time series trajectories is an important problem in many domains such as medicine, economics, and vision. It is often unclear what is the appropriate metric to use because of the complex nature of noise in the trajectories (e.g. different sampling rates or outliers). Experts typically hand-craft or manually select a specific metric, such as Dynamic Time Warping (DTW), to apply on their data. In this paper, we propose an end-to-end framework, autowarp, that optimizes and learns a good metric given unlabeled trajectories. We define a flexible and differentiable family of warping metrics, which encompasses common metrics such as DTW, Edit Distance, Euclidean, etc. Autowarp then leverages the representation power of sequence autoencoders to optimize for a member of this warping family. The output is an metric which is easy to interpret and can be robustly learned from relatively few trajectories. In systematic experiments across different domains, we show that autowarp often outperforms hand-crafted trajectory similarity metrics.

Precision and Recall for Time Series

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #116

Nesime Tatbul · Tae Jun Lee · Stan Zdonik · Mejbah Alam · Justin Gottschlich

Classical anomaly detection is principally concerned with point-based anomalies, those anomalies that occur at a single point in time. Yet, many real-world anomalies are range-based, meaning they occur over a period of time. Motivated by this observation, we present a new mathematical model to evaluate the accuracy of time series classification algorithms. Our model expands the well-known Precision and Recall metrics to measure ranges, while simultaneously enabling customization support for domain-specific preferences.

Deep Generative Markov State Models

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #117

Hao Wu · Andreas Mardt · Luca Pasquali · Frank Noe

We propose a deep generative Markov State Model (DeepGenMSM) learning framework for inference of metastable dynamical systems and prediction of trajectories. After unsupervised training on time series data, the model contains (i) a probabilistic encoder that maps from high-dimensional configuration space to a small-sized vector indicating the membership to metastable (long-lived) states, (ii) a Markov chain that governs the transitions between metastable states and facilitates analysis of the long-time dynamics, and (iii) a generative part that samples the conditional distribution of configurations in the next time step. The model can be operated in a recursive fashion to generate trajectories to predict the system evolution from a defined starting state and propose new configurations. The DeepGenMSM is demonstrated to provide accurate estimates of the long-time kinetics and generate valid distributions for molecular dynamics (MD) benchmark systems. Remarkably, we show that DeepGenMSMs are able to make long time-steps in molecular configuration space and generate physically realistic structures in regions that were not seen in training data.

Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-Divergences

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #118

Jeremias Knoblauch · Jack E Jewson · Theodoros Damoulas

We present the very first robust Bayesian Online Changepoint Detection algorithm through General Bayesian Inference (GBI) with β-divergences. The resulting inference procedure is doubly robust for both the predictive and the changepoint (CP) posterior, with linear time and constant space complexity. We provide a construction for exponential models and demonstrate it on the Bayesian Linear Regression model. In so doing, we make two additional contributions: Firstly, we make GBI scalable using Structural Variational approximations that are exact as $\beta \to 0$. Secondly, we give a principled way of choosing the divergence parameter β by minimizing expected predictive loss on-line. Reducing False Discovery Rates of \CPs from up to 99\% to 0\% on real world data, this offers the state of the art.

Regularization Learning Networks: Deep Learning for Tabular Datasets

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #119

Ira Shavitt · Eran Segal

Despite their impressive performance, Deep Neural Networks (DNNs) typically underperform Gradient Boosting Trees (GBTs) on many tabular-dataset learning tasks. We propose that applying a different regularization coefficient to each weight might boost the performance of DNNs by allowing them to make more use of the more relevant inputs. However, this will lead to an intractable number of hyperparameters. Here, we introduce Regularization Learning Networks (RLNs), which overcome this challenge by introducing an efficient hyperparameter tuning scheme which minimizes a new Counterfactual Loss. Our results show that RLNs significantly improve DNNs on tabular datasets, and achieve comparable results to GBTs, with the best performance achieved with an ensemble that combines GBTs and RLNs. RLNs produce extremely sparse networks, eliminating up to 99.8% of the network edges and 82% of the input features, thus providing more interpretable models and reveal the importance that the network assigns to different inputs. RLNs could efficiently learn a single network in datasets that comprise both tabular and unstructured data, such as in the setting of medical imaging accompanied by electronic health records. An open source implementation of RLN can be found at https://github.com/irashavitt/regularizationlearningnetworks.

Generative modeling for protein structures

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #120

Namrata Anand · Possu Huang

Analyzing the structure and function of proteins is a key part of understanding biology at the molecular and cellular level. In addition, a major engineering challenge is to design new proteins in a principled and methodical way. Current computational modeling methods for protein design are slow and often require human oversight and intervention. Here, we apply Generative Adversarial Networks (GANs) to the task of generating protein structures, toward application in fast de novo protein design. We encode protein structures in terms of pairwise distances between alpha-carbons on the protein backbone, which eliminates the need for the generative model to learn translational and rotational symmetries. We then introduce a convex formulation of corruption-robust 3D structure recovery to fold the protein structures from generated pairwise distance maps, and solve these problems using the Alternating Direction Method of Multipliers. We test the effectiveness of our models by predicting completions of corrupted protein structures and show that the method is capable of quickly producing structurally plausible solutions.

Geometry Based Data Generation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #121

Ofir Lindenbaum · Jay Stanley · Guy Wolf · Smita Krishnaswamy

We propose a new type of generative model for high-dimensional data that learns a manifold geometry of the data, rather than density, and can generate points evenly along this manifold. This is in contrast to existing generative models that represent data density, and are strongly affected by noise and other artifacts of data collection. We demonstrate how this approach corrects sampling biases and artifacts, thus improves several downstream data analysis tasks, such as clustering and classification. Finally, we demonstrate that this approach is especially useful in biology where, despite the advent of single-cell technologies, rare subpopulations and gene-interaction relationships are affected by biased sampling. We show that SUGAR can generate hypothetical populations, and it is able to reveal intrinsic patterns and mutual-information relationships between genes on a single-cell RNA sequencing dataset of hematopoiesis.

Learning Concave Conditional Likelihood Models for Improved Analysis of Tandem Mass Spectra

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #122

John T Halloran · David M Rocke

The most widely used technology to identify the proteins present in a complex biological sample is tandem mass spectrometry, which quickly produces a large collection of spectra representative of the peptides (i.e., protein subsequences) present in the original sample. In this work, we greatly expand the parameter learning capabilities of a dynamic Bayesian network (DBN) peptide-scoring algorithm, Didea, by deriving emission distributions for which its conditional log-likelihood scoring function remains concave. We show that this class of emission distributions, called Convex Virtual Emissions (CVEs), naturally generalizes the log-sum-exp function while rendering both maximum likelihood estimation and conditional maximum likelihood estimation concave for a wide range of Bayesian networks. Utilizing CVEs in Didea allows efficient learning of a large number of parameters while ensuring global convergence, in stark contrast to Didea’s previous parameter learning framework (which could only learn a single parameter using a costly grid search) and other trainable models (which only ensure convergence to local optima). The newly trained scoring function substantially outperforms the state-of-the-art in both scoring function accuracy and downstream Fisher kernel analysis. Furthermore, we significantly improve Didea’s runtime performance through successive optimizations to its message passing schedule and derive explicit connections between Didea’s new concave score and related MS/MS scoring functions.

Generalizing Tree Probability Estimation via Bayesian Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #123

Cheng Zhang · Frederick A Matsen IV

Probability estimation is one of the fundamental tasks in statistics and machine learning. However, standard methods for probability estimation on discrete objects do not handle object structure in a satisfactory manner. In this paper, we derive a general Bayesian network formulation for probability estimation on leaf-labeled trees that enables flexible approximations which can generalize beyond observations. We show that efficient algorithms for learning Bayesian networks can be easily extended to probability estimation on this challenging structured space. Experiments on both synthetic and real data show that our methods greatly outperform the current practice of using the empirical distribution, as well as a previous effort for probability estimation on trees.

Learning Temporal Point Processes via Reinforcement Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #124

Shuang Li · Shuai Xiao · Shixiang Zhu · Nan Du · Yao Xie · Le Song

Social goods, such as healthcare, smart city, and information networks, often produce ordered event data in continuous time. The generative processes of these event data can be very complex, requiring flexible models to capture their dynamics. Temporal point processes offer an elegant framework for modeling event data without discretizing the time. However, the existing maximum-likelihood-estimation (MLE) learning paradigm requires hand-crafting the intensity function beforehand and cannot directly monitor the goodness-of-fit of the estimated model in the process of training. To alleviate the risk of model-misspecification in MLE, we propose to generate samples from the generative model and monitor the quality of the samples in the process of training until the samples and the real data are indistinguishable. We take inspiration from reinforcement learning (RL) and treat the generation of each event as the action taken by a stochastic policy. We parameterize the policy as a flexible recurrent neural network and gradually improve the policy to mimic the observed event distribution. Since the reward function is unknown in this setting, we uncover an analytic and nonparametric form of the reward function using an inverse reinforcement learning formulation. This new RL framework allows us to derive an efficient policy gradient algorithm for learning flexible point process models, and we show that it performs well in both synthetic and real data.

Exponentially Weighted Imitation Learning for Batched Historical Data

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #125

Qing Wang · Jiechao Xiong · Lei Han · peng sun · Han Liu · Tong Zhang

We consider deep policy learning with only batched historical trajectories. The main challenge of this problem is that the learner no longer has a simulator or ``environment oracle'' as in most reinforcement learning settings. To solve this problem, we propose a monotonic advantage reweighted imitation learning strategy that is applicable to problems with complex nonlinear function approximation and works well with hybrid (discrete and continuous) action space. The method does not rely on the knowledge of the behavior policy, thus can be used to learn from data generated by an unknown policy. Under mild conditions, our algorithm, though surprisingly simple, has a policy improvement bound and outperforms most competing methods empirically. Thorough numerical results are also provided to demonstrate the efficacy of the proposed methodology.

Evidential Deep Learning to Quantify Classification Uncertainty

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #126

Murat Sensoy · Lance Kaplan · Melih Kandemir

Deterministic neural nets have been shown to learn effective predictors on a wide range of machine learning problems. However, as the standard approach is to train the network to minimize a prediction loss, the resultant model remains ignorant to its prediction confidence. Orthogonally to Bayesian neural nets that indirectly infer prediction uncertainty through weight uncertainties, we propose explicit modeling of the same using the theory of subjective logic. By placing a Dirichlet distribution on the class probabilities, we treat predictions of a neural net as subjective opinions and learn the function that collects the evidence leading to these opinions by a deterministic neural net from data. The resultant predictor for a multi-class classification problem is another Dirichlet distribution whose parameters are set by the continuous output of a neural net. We provide a preliminary analysis on how the peculiarities of our new loss function drive improved uncertainty estimation. We observe that our method achieves unprecedented success on detection of out-of-distribution queries and endurance against adversarial perturbations.

Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #127

Amit Dhurandhar · Pin-Yu Chen · Ronny Luss · Chun-Chen Tu · Paishun Ting · Karthikeyan Shanmugam · Payel Das

In this paper we propose a novel method that provides contrastive explanations justifying the classification of an input by a black box classifier such as a deep neural network. Given an input we find what should be minimally and sufficiently present (viz. important object pixels in an image) to justify its classification and analogously what should be minimally and necessarily \emph{absent} (viz. certain background pixels). We argue that such explanations are natural for humans and are used commonly in domains such as health care and criminology. What is minimally but critically \emph{absent} is an important part of an explanation, which to the best of our knowledge, has not been explicitly identified by current explanation methods that explain predictions of neural networks. We validate our approach on three real datasets obtained from diverse domains; namely, a handwritten digits dataset MNIST, a large procurement fraud dataset and a brain activity strength dataset. In all three cases, we witness the power of our approach in generating precise explanations that are also easy for human experts to understand and evaluate.

Towards Robust Interpretability with Self-Explaining Neural Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #128

David Alvarez Melis · Tommi Jaakkola

Most recent work on interpretability of complex machine learning models has focused on estimating a-posteriori explanations for previously trained models around specific predictions. Self-explaining models where interpretability plays a key role already during learning have received much less attention. We propose three desiderata for explanations in general -- explicitness, faithfulness, and stability -- and show that existing methods do not satisfy them. In response, we design self-explaining models in stages, progressively generalizing linear classifiers to complex yet architecturally explicit models. Faithfulness and stability are enforced via regularization specifically tailored to such models. Experimental results across various benchmark datasets show that our framework offers a promising direction for reconciling model complexity and interpretability.

Model Agnostic Supervised Local Explanations

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #129

Gregory Plumb · Denali Molitor · Ameet Talwalkar

Model interpretability is an increasingly important component of practical machine learning. Some of the most common forms of interpretability systems are example-based, local, and global explanations. One of the main challenges in interpretability is designing explanation systems that can capture aspects of each of these explanation types, in order to develop a more thorough understanding of the model. We address this challenge in a novel model called MAPLE that uses local linear modeling techniques along with a dual interpretation of random forests (both as a supervised neighborhood approach and as a feature selection method). MAPLE has two fundamental advantages over existing interpretability systems. First, while it is effective as a black-box explanation system, MAPLE itself is a highly accurate predictive model that provides faithful self explanations, and thus sidesteps the typical accuracy-interpretability trade-off. Specifically, we demonstrate, on several UCI datasets, that MAPLE is at least as accurate as random forests and that it produces more faithful local explanations than LIME, a popular interpretability system. Second, MAPLE provides both example-based and local explanations and can detect global patterns, which allows it to diagnose limitations in its local explanations.

To Trust Or Not To Trust A Classifier

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #130

Heinrich Jiang · Been Kim · Melody Guan · Maya Gupta

Knowing when a classifier's prediction can be trusted is useful in many applications and critical for safely using AI. While the bulk of the effort in machine learning research has been towards improving classifier performance, understanding when a classifier's predictions should and should not be trusted has received far less attention. The standard approach is to use the classifier's discriminant or confidence score; however, we show there exists an alternative that is more effective in many situations. We propose a new score, called the {\it trust score}, which measures the agreement between the classifier and a modified nearest-neighbor classifier on the testing example. We show empirically that high (low) trust scores produce surprisingly high precision at identifying correctly (incorrectly) classified examples, consistently outperforming the classifier's confidence score as well as many other baselines. Further, under some mild distributional assumptions, we show that if the trust score for an example is high (low), the classifier will likely agree (disagree) with the Bayes-optimal classifier. Our guarantees consist of non-asymptotic rates of statistical consistency under various nonparametric settings and build on recent developments in topological data analysis.

Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #131

David Madras · Toni Pitassi · Richard Zemel

In many machine learning applications, there are multiple decision-makers involved, both automated and human. The interaction between these agents often goes unaddressed in algorithmic development. In this work, we explore a simple version of this interaction with a two-stage framework containing an automated model and an external decision-maker. The model can choose to say PASS, and pass the decision downstream, as explored in rejection learning. We extend this concept by proposing "learning to defer", which generalizes rejection learning by considering the effect of other agents in the decision-making process. We propose a learning algorithm which accounts for potential biases held by external decision-makers in a system. Experiments demonstrate that learning to defer can make systems not only more accurate but also less biased. Even when working with inconsistent or biased users, we show that deferring models still greatly improve the accuracy and/or fairness of the entire system.

Hunting for Discriminatory Proxies in Linear Regression Models

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #132

Samuel Yeom · Anupam Datta · Matt Fredrikson

A machine learning model may exhibit discrimination when used to make decisions involving people. One potential cause for such outcomes is that the model uses a statistical proxy for a protected demographic attribute. In this paper we formulate a definition of proxy use for the setting of linear regression and present algorithms for detecting proxies. Our definition follows recent work on proxies in classification models, and characterizes a model's constituent behavior that: 1) correlates closely with a protected random variable, and 2) is causally influential in the overall behavior of the model. We show that proxies in linear regression models can be efficiently identified by solving a second-order cone program, and further extend this result to account for situations where the use of a certain input variable is justified as a ``business necessity''. Finally, we present empirical results on two law enforcement datasets that exhibit varying degrees of racial disparity in prediction outcomes, demonstrating that proxies shed useful light on the causes of discriminatory behavior in models.

Empirical Risk Minimization Under Fairness Constraints

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #133

Michele Donini · Luca Oneto · Shai Ben-David · John Shawe-Taylor · Massimiliano Pontil

We address the problem of algorithmic fairness: ensuring that sensitive information does not unfairly influence the outcome of a classifier. We present an approach based on empirical risk minimization, which incorporates a fairness constraint into the learning problem. It encourages the conditional risk of the learned classifier to be approximately constant with respect to the sensitive variable. We derive both risk and fairness bounds that support the statistical consistency of our methodology. We specify our approach to kernel methods and observe that the fairness requirement implies an orthogonality constraint which can be easily added to these methods. We further observe that for linear models the constraint translates into a simple data preprocessing step. Experiments indicate that the method is empirically effective and performs favorably against state-of-the-art approaches.

Approximation algorithms for stochastic clustering

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #134

David Harris · Shi Li · Aravind Srinivasan · Khoa Trinh · Thomas Pensyl

We consider stochastic settings for clustering, and develop provably-good (approximation) algorithms for a number of these notions. These algorithms allow one to obtain better approximation ratios compared to the usual deterministic clustering setting. Additionally, they offer a number of advantages including providing fairer clustering and clustering which has better long-term behavior for each user. In particular, they ensure that every user is guaranteed to get good service (on average). We also complement some of these with impossibility results.

Re-evaluating evaluation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #135

David Balduzzi · Karl Tuyls · Julien Perolat · Thore Graepel

Progress in machine learning is measured by careful evaluation on problems of outstanding common interest. However, the proliferation of benchmark suites and environments, adversarial attacks, and other complications has diluted the basic evaluation model by overwhelming researchers with choices. Deliberate or accidental cherry picking is increasingly likely, and designing well-balanced evaluation suites requires increasing effort. In this paper we take a step back and propose Nash averaging. The approach builds on a detailed analysis of the algebraic structure of evaluation in two basic scenarios: agent-vs-agent and agent-vs-task. The key strength of Nash averaging is that it automatically adapts to redundancies in evaluation data, so that results are not biased by the incorporation of easy tasks or weak agents. Nash averaging thus encourages maximally inclusive evaluation -- since there is no harm (computational cost aside) from including all available tasks and agents.

Does mitigating ML's impact disparity require treatment disparity?

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #136

Zachary Lipton · Julian McAuley · Alexandra Chouldechova

Following precedent in employment discrimination law, two notions of disparity are widely-discussed in papers on fairness and ML. Algorithms exhibit treatment disparity if they formally treat members of protected subgroups differently; algorithms exhibit impact disparity when outcomes differ across

subgroups (even unintentionally). Naturally, we can achieve impact parity through purposeful treatment disparity. One line of papers aims to reconcile the two parities proposing disparate learning processes (DLPs). Here, the sensitive feature is used during training but a group-blind classifier is produced. In this paper, we show that: (i) when sensitive and (nominally) nonsensitive features are correlated, DLPs will indirectly implement treatment disparity, undermining the policy desiderata they are designed to address; (ii) when group membership is partly revealed by other features, DLPs induce within-class discrimination; and (iii) in general, DLPs provide suboptimal trade-offs between accuracy and impact parity. Experimental results on several real-world datasets highlight the practical consequences of applying DLPs.

Enhancing the Accuracy and Fairness of Human Decision Making

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #137

Isabel Valera · Adish Singla · Manuel Gomez Rodriguez

Societies often rely on human experts to take a wide variety of decisions affecting their members, from jail-or-release decisions taken by judges and stop-and-frisk decisions taken by police officers to accept-or-reject decisions taken by academics. In this context, each decision is taken by an expert who is typically chosen uniformly at random from a pool of experts. However, these decisions may be imperfect due to limited experience, implicit biases, or faulty probabilistic reasoning. Can we improve the accuracy and fairness of the overall decision making process by optimizing the assignment between experts and decisions? In this paper, we address the above problem from the perspective of sequential decision making and show that, for different fairness notions from the literature, it reduces to a sequence of (constrained) weighted bipartite matchings, which can be solved efficiently using algorithms with approximation guarantees. Moreover, these algorithms also benefit from posterior sampling to actively trade off exploitation---selecting expert assignments which lead to accurate and fair decisions---and exploration---selecting expert assignments to learn about the experts' preferences and biases. We demonstrate the effectiveness of our algorithms on both synthetic and real-world data and show that they can significantly improve both the accuracy and fairness of the decisions taken by pools of experts.

The Price of Fair PCA: One Extra dimension

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #138

Samira Samadi · Uthaipon Tantipongpipat · Jamie Morgenstern · Mohit Singh · Santosh Vempala

We investigate whether the standard dimensionality reduction technique of PCA inadvertently produces data representations with different fidelity for two different populations. We show on

several real-world data sets, PCA has higher reconstruction error on population A than on B (for example, women versus men or lower- versus higher-educated individuals). This can happen even when the data set has a similar number of samples from A and B. This motivates our study of dimensionality reduction techniques which maintain similar fidelity for A and B. We define the notion of Fair PCA and give a polynomial-time algorithm for finding a low dimensional representation of the data which is nearly-optimal with respect to this measure. Finally, we show on real-world data sets that our algorithm can be used to efficiently generate a fair low dimensional representation of the data.

Practical Methods for Graph Two-Sample Testing

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #139

Debarghya Ghoshdastidar · Ulrike von Luxburg

Hypothesis testing for graphs has been an important tool in applied research fields for more than two decades, and still remains a challenging problem as one often needs to draw inference from few replicates of large graphs. Recent studies in statistics and learning theory have provided some theoretical insights about such high-dimensional graph testing problems, but the practicality of the developed theoretical methods remains an open question. In this paper, we consider the problem of two-sample testing of large graphs. We demonstrate the practical merits and limitations of existing theoretical tests and their bootstrapped variants. We also propose two new tests based on asymptotic distributions. We show that these tests are computationally less expensive and, in some cases, more reliable than the existing methods.

Topkapi: Parallel and Fast Sketches for Finding Top-K Frequent Elements

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #140

Ankush Mandal · He Jiang · Anshumali Shrivastava · Vivek Sarkar

Identifying the top-K frequent items is one of the most common and important operations in large data processing systems. As a result, several solutions have been proposed to solve this problem approximately. In this paper, we identify that in modern distributed settings with both multi-node as well as multi-core parallelism, existing algorithms, although theoretically sound, are suboptimal from the performance perspective. In particular, for identifying top-K frequent items, Count-Min Sketch (CMS) has fantastic update time but lack the important property of reducibility which is needed for exploiting available massive data parallelism. On the other end, popular Frequent algorithm (FA) leads to reducible summaries but the update costs are significant. In this paper, we present Topkapi, a fast and parallel algorithm for finding top-K frequent items, which gives the best of both worlds, i.e., it is reducible as well as efficient update time similar to CMS. Topkapi possesses strong theoretical guarantees and leads to significant performance gains due to increased parallelism, relative to past work.

KDGAN: Knowledge Distillation with Generative Adversarial Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #141

Xiaojie Wang · Rui Zhang · Yu Sun · Jianzhong Qi

Knowledge distillation (KD) aims to train a lightweight classifier suitable to provide accurate inference with constrained resources in multi-label learning. Instead of directly consuming feature-label pairs, the classifier is trained by a teacher, i.e., a high-capacity model whose training may be resource-hungry. The accuracy of the classifier trained this way is usually suboptimal because it is difficult to learn the true data distribution from the teacher. An alternative method is to adversarially train the classifier against a discriminator in a two-player game akin to generative adversarial networks (GAN), which can ensure the classifier to learn the true data distribution at the equilibrium of this game. However, it may take excessively long time for such a two-player game to reach equilibrium due to high-variance gradient updates. To address these limitations, we propose a three-player game named KDGAN consisting of a classifier, a teacher, and a discriminator. The classifier and the teacher learn from each other via distillation losses and are adversarially trained against the discriminator via adversarial losses. By simultaneously optimizing the distillation and adversarial losses, the classifier will learn the true data distribution at the equilibrium. We approximate the discrete distribution learned by the classifier (or the teacher) with a concrete distribution. From the concrete distribution, we generate continuous samples to obtain low-variance gradient updates, which speed up the training. Extensive experiments using real datasets confirm the superiority of KDGAN in both accuracy and training speed.

Modeling Dynamic Missingness of Implicit Feedback for Recommendation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #142

Menghan Wang · Mingming Gong · Xiaolin Zheng · Kun Zhang

Implicit feedback is widely used in collaborative filtering methods for recommendation. It is well known that implicit feedback contains a large number of values that are \emph{missing not at random} (MNAR); and the missing data is a mixture of negative and unknown feedback, making it difficult to learn user's negative preferences. Recent studies modeled \emph{exposure}, a latent missingness variable which indicates whether an item is missing to a user, to give each missing entry a confidence of being negative feedback. However, these studies use static models and ignore the information in temporal dependencies among items, which seems to be a essential underlying factor to subsequent missingness. To model and exploit the dynamics of missingness, we propose a latent variable named ``\emph{user intent}'' to govern the temporal changes of item missingness, and a hidden Markov model to represent such a process. The resulting framework captures the dynamic item missingness and incorporate it into matrix factorization (MF) for recommendation. We also explore two types of constraints to achieve a more compact and interpretable representation of \emph{user intents}. Experiments on real-world datasets demonstrate the superiority of our method against state-of-the-art recommender systems.

Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #143

Trong Dinh Thac Do · Longbing Cao

A conjugate Gamma-Poisson model for Dynamic Matrix Factorization incorporated with metadata influence (mGDMF for short) is proposed to effectively and efficiently model massive, sparse and dynamic data in recommendations. Modeling recommendation problems with a massive number of ratings and very sparse or even no ratings on some users/items in a dynamic setting is very demanding and poses critical challenges to well-studied matrix factorization models due to the large-scale, sparse and dynamic nature of the data. Our proposed mGDMF tackles these challenges by introducing three strategies: (1) constructing a stable Gamma-Markov chain model that smoothly drifts over time by combining both static and dynamic latent features of data; (2) incorporating the user/item metadata into the model to tackle sparse ratings; and (3) undertaking stochastic variational inference to efficiently handle massive data. mGDMF is conjugate, dynamic and scalable. Experiments show that mGDMF significantly (both effectively and efficiently) outperforms the state-of-the-art static and dynamic models on large, sparse and dynamic data.

Non-metric Similarity Graphs for Maximum Inner Product Search

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #144

Stanislav Morozov · Artem Babenko

In this paper we address the problem of Maximum Inner Product Search (MIPS) that is currently the computational bottleneck in a large number of machine learning applications. While being similar to the nearest neighbor search (NNS), the MIPS problem was shown to be more challenging, as the inner product is not a proper metric function. We propose to solve the MIPS problem with the usage of similarity graphs, i.e., graphs where each vertex is connected to the vertices that are the most similar in terms of some similarity function. Originally, the framework of similarity graphs was proposed for metric spaces and in this paper we naturally extend it to the non-metric MIPS scenario. We demonstrate that, unlike existing approaches, similarity graphs do not require any data transformation to reduce MIPS to the NNS problem and should be used for the original data. Moreover, we explain why such a reduction is detrimental for similarity graphs. By an extensive comparison to the existing approaches, we show that the proposed method is a game-changer in terms of the runtime/accuracy trade-off for the MIPS problem.

Norm-Ranging LSH for Maximum Inner Product Search

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #145

Xiao Yan · Jinfeng Li · Xinyan Dai · Hongzhi Chen · James Cheng

Neyshabur and Srebro proposed SIMPLE-LSH, which is the state-of-the-art hashing based algorithm for maximum inner product search (MIPS). We found that the performance of SIMPLE-LSH, in both theory and practice, suffers from long tails in the 2-norm distribution of real datasets. We propose NORM-RANGING LSH, which addresses the excessive normalization problem caused by long tails by partitioning a dataset into sub-datasets and building a hash index for each sub-dataset independently. We prove that NORM-RANGING LSH achieves lower query time complexity than SIMPLE-LSH under mild conditions. We also show that the idea of dataset partitioning can improve another hashing based MIPS algorithm. Experiments show that NORM-RANGING LSH probes much less items than SIMPLE-LSH at the same recall, thus significantly benefiting MIPS based applications.

A Dual Framework for Low-rank Tensor Completion

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #146

Madhav Nimishakavi · Pratik Kumar Jawanpuria · Bamdev Mishra

One of the popular approaches for low-rank tensor completion is to use the latent trace norm regularization. However, most existing works in this direction learn a sparse combination of tensors. In this work, we fill this gap by proposing a variant of the latent trace norm that helps in learning a non-sparse combination of tensors. We develop a dual framework for solving the low-rank tensor completion problem. We first show a novel characterization of the dual solution space with an interesting factorization of the optimal solution. Overall, the optimal solution is shown to lie on a Cartesian product of Riemannian manifolds. Furthermore, we exploit the versatile Riemannian optimization framework for proposing computationally efficient trust region algorithm. The experiments illustrate the efficacy of the proposed algorithm on several real-world datasets across applications.

Low-Rank Tucker Decomposition of Large Tensors Using TensorSketch

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #147

Osman Asif Malik · Stephen Becker

We propose two randomized algorithms for low-rank Tucker decomposition of tensors. The algorithms, which incorporate sketching, only require a single pass of the input tensor and can handle tensors whose elements are streamed in any order. To the best of our knowledge, ours are the only algorithms which can do this. We test our algorithms on sparse synthetic data and compare them to multiple other methods. We also apply one of our algorithms to a real dense 38 GB tensor representing a video and use the resulting decomposition to correctly classify frames containing disturbances.

Dropping Symmetry for Fast Symmetric Nonnegative Matrix Factorization

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #148

Zhihui Zhu · Xiao Li · Kai Liu · Qiuwei Li

Symmetric nonnegative matrix factorization (NMF)---a special but important class of the general NMF---is demonstrated to be useful for data analysis and in particular for various clustering tasks. Unfortunately, designing fast algorithms for Symmetric NMF is not as easy as for the nonsymmetric counterpart, the latter admitting the splitting property that allows efficient alternating-type algorithms. To overcome this issue, we transfer the symmetric NMF to a nonsymmetric one, then we can adopt the idea from the state-of-the-art algorithms for nonsymmetric NMF to design fast algorithms solving symmetric NMF. We rigorously establish that solving nonsymmetric reformulation returns a solution for symmetric NMF and then apply fast alternating based algorithms for the corresponding reformulated problem. Furthermore, we show these fast algorithms admit strong convergence guarantee in the sense that the generated sequence is convergent at least at a sublinear rate and it converges globally to a critical point of the symmetric NMF. We conduct experiments on both synthetic data and image clustering to support our result.

Semidefinite relaxations for certifying robustness to adversarial examples

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #149

Aditi Raghunathan · Jacob Steinhardt · Percy Liang

Despite their impressive performance on diverse tasks, neural networks fail catastrophically in the presence of adversarial inputs—imperceptibly but adversarially perturbed versions of natural inputs. We have witnessed an arms race between defenders who attempt to train robust networks and attackers who try to construct adversarial examples. One promise of ending the arms race is developing certified defenses, ones which are provably robust against all attackers in some family. These certified defenses are based on convex relaxations which construct an upper bound on the worst case loss over all attackers in the family. Previous relaxations are loose on networks that are not trained against the respective relaxation. In this paper, we propose a new semidefinite relaxation for certifying robustness that applies to arbitrary ReLU networks. We show that our proposed relaxation is tighter than previous relaxations and produces meaningful robustness guarantees on three different foreign networks whose training objectives are agnostic to our proposed relaxation.

Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #150

Borja Balle · Gilles Barthe · Marco Gaboardi

Differential privacy comes equipped with multiple analytical tools for the design of private data analyses. One important tool is the so-called "privacy amplification by subsampling" principle, which ensures that a differentially private mechanism run on a random subsample of a population provides higher privacy guarantees than when run on the entire population. Several instances of this principle have been studied for different random subsampling methods, each with an ad-hoc analysis. In this paper we present a general method that recovers and improves prior analyses, yields lower bounds and derives new instances of privacy amplification by subsampling. Our method leverages a characterization of differential privacy as a divergence which emerged in the program verification community. Furthermore, it introduces new tools, including advanced joint convexity and privacy profiles, which might be of independent interest.

Differentially Private Testing of Identity and Closeness of

Discrete Distributions

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #151

Jayadev Acharya · Ziteng Sun · Huanyu Zhang

We study the fundamental problems of identity testing (goodness of fit), and closeness testing (two sample test) of distributions over k elements, under differential privacy. While the problems have a long history in statistics, finite sample bounds for these problems have only been established recently.

In this work, we derive upper and lower bounds on the sample complexity of both the problems under (ε, δ)-differential privacy. We provide optimal sample complexity algorithms for identity testing problem for all parameter ranges, and the first results for closeness testing. Our closeness testing bounds are optimal in the sparse regime where the number of samples is at most k.

Our upper bounds are obtained by privatizing non-private estimators for these problems. The non-private estimators are chosen to have small sensitivity. We propose a general framework to establish lower bounds on the sample complexity of statistical tasks under differential privacy. We show a bound on differentially private algorithms in terms of a coupling between the two hypothesis classes we aim to test. By constructing carefully chosen priors over the hypothesis classes, and using Le Cam's two point theorem we provide a general mechanism for proving lower bounds. We believe that the framework can be used to obtain strong lower bounds for other statistical tasks under privacy.

Differentially Private k-Means with Constant Multiplicative Error

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #152

Uri Stemmer · Haim Kaplan

We design new differentially private algorithms for the Euclidean k-means problem, both in the centralized model and in the local model of differential privacy. In both models, our algorithms achieve significantly improved error guarantees than the previous state-of-the-art. In addition, in the local model, our algorithm significantly reduces the number of interaction rounds. Although the problem has been widely studied in the context of differential privacy, all of the existing constructions achieve only super constant approximation factors. We present, for the first time, efficient private algorithms for the problem with constant multiplicative error. Furthermore, we show how to modify our algorithms so they compute private coresets for k-means clustering in both models.

Local Differential Privacy for Evolving Data

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #153

Matthew Joseph · Aaron Roth · Jonathan Ullman · Bo Waggoner

There are now several large scale deployments of differential privacy used to collect statistical information about users. However, these deployments periodically recollect the data and recompute the statistics using algorithms designed for a single use. As a result, these systems do not provide meaningful privacy guarantees over long time scales. Moreover, existing techniques to mitigate this effect do not apply in the ``local model'' of differential privacy that these systems use. In this paper, we introduce a new technique for local differential privacy that makes it possible to maintain up-to-date statistics over time, with privacy guarantees that degrade only in the number of changes in the underlying distribution rather than the number of collection periods. We use our technique for tracking a changing statistic in the setting where users are partitioned into an unknown collection of groups, and at every time period each user draws a single bit from a common (but changing) group-specific distribution. We also provide an application to frequency and heavy-hitter estimation.

Adversarial Attacks on Stochastic Bandits

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #154

Kwang-Sung Jun · Lihong Li · Yuzhe Ma · Jerry Zhu

We study adversarial attacks that manipulate the reward signals to control the actions chosen by a stochastic multi-armed bandit algorithm. We propose the first attack against two popular bandit algorithms: ϵ-greedy and UCB, \emph{without} knowledge of the mean rewards. The attacker is able to spend only logarithmic effort, multiplied by a problem-specific parameter that becomes smaller as the bandit problem gets easier to attack. The result means the attacker can easily hijack the behavior of the bandit algorithm to promote or obstruct certain actions, say, a particular medical treatment. As bandits are seeing increasingly wide use in practice, our study exposes a significant security threat.

Distributed Learning without Distress: Privacy-Preserving Empirical Risk Minimization

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #155

Bargav Jayaraman · Lingxiao Wang · David Evans · Quanquan Gu

Distributed learning allows a group of independent data owners to collaboratively learn a model over their data sets without exposing their private data. We present a distributed learning approach that combines differential privacy with secure multi-party computation. We explore two popular methods of differential privacy, output perturbation and gradient perturbation, and advance the state-of-the-art for both methods in the distributed learning setting. In our output perturbation method, the parties combine local models within a secure computation and then add the required differential privacy noise before revealing the model. In our gradient perturbation method, the data owners collaboratively train a global model via an iterative learning algorithm. At each iteration, the parties aggregate their local gradients within a secure computation, adding sufficient noise to ensure privacy before the gradient updates are revealed. For both methods, we show that the noise can be reduced in the multi-party setting by adding the noise inside the secure computation after aggregation, asymptotically improving upon the best previous results. Experiments on real world data sets demonstrate that our methods provide substantial utility gains for typical privacy requirements.

A Spectral View of Adversarially Robust Features

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #156

Shivam Garg · Vatsal Sharan · Brian Zhang · Gregory Valiant

Given the apparent difficulty of learning models that are robust to adversarial perturbations, we propose tackling the simpler problem of developing adversarially robust features. Specifically, given a dataset and metric of interest, the goal is to return a function (or multiple functions) that 1) is robust to adversarial perturbations, and 2) has significant variation across the datapoints. We establish strong connections between adversarially robust features and a natural spectral property of the geometry of the dataset and metric of interest. This connection can be leveraged to provide both robust features, and a lower bound on the robustness of any function that has significant variance across the dataset. Finally, we provide empirical evidence that the adversarially robust features given by this spectral approach can be fruitfully leveraged to learn a robust (and accurate) model.

Efficient Formal Safety Analysis of Neural Networks

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #157

Shiqi Wang · Kexin Pei · Justin Whitehouse · Junfeng Yang · Suman Jana

Neural networks are increasingly deployed in real-world safety-critical domains such as autonomous driving, aircraft collision avoidance, and malware detection. However, these networks have been shown to often mispredict on inputs with minor adversarial or even accidental perturbations.

Consequences of such errors can be disastrous and even potentially fatal as shown by the recent Tesla autopilot crash. Thus, there is an urgent need for formal analysis systems that can rigorously check neural networks for violations of different safety properties such as robustness against adversarial perturbations within a certain L-norm of a given image. An effective safety analysis system for a neural network must be able to either ensure that a safety property is satisfied by the network or find a counterexample, i.e., an input for which the network will violate the property. Unfortunately, most existing techniques for performing such analysis struggle to scale beyond very small networks and the ones that can scale to larger networks suffer from high false positives and cannot produce concrete counterexamples in case of a property violation. In this paper, we present a new efficient approach for rigorously checking different safety properties of neural networks that significantly outperforms existing approaches by multiple orders of magnitude. Our approach can check different safety properties and find concrete counterexamples for networks that are 10x larger than the ones supported by existing analysis techniques. We believe that our approach to estimating tight output bounds of a network for a given input range can also help improve the explainability of neural networks and guide the training process of more robust neural networks.

Contamination Attacks and Mitigation in Multi-Party Machine Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #158

Jamie Hayes · Olga Ohrimenko

Machine learning is data hungry; the more data a model has access to in training, the more likely it is to perform well at inference time. Distinct parties may want to combine their local data to gain the benefits of a model trained on a large corpus of data. We consider such a case: parties get access to the model trained on their joint data but do not see each others individual datasets. We show that one needs to be careful when using this multi-party model since a potentially malicious party can taint the model by providing contaminated data. We then show how adversarial training can defend against such attacks by preventing the model from learning trends specific to individual parties data, thereby also guaranteeing party-level membership privacy.

Explaining Deep Learning Models -- A Bayesian Non-parametric Approach

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #159

Wenbo Guo · Sui Huang · Yunzhe Tao · Xinyu Xing · Lin Lin

Understanding and interpreting how machine learning (ML) models make decisions have been a big challenge. While recent research has proposed various technical approaches to provide some clues as to how an ML model makes individual predictions, they cannot provide users with an ability to inspect a model as a complete entity. In this work, we propose a novel technical approach that augments a Bayesian non-parametric regression mixture model with multiple elastic nets. Using the enhanced mixture model, we can extract generalizable insights for a target model through a global approximation. To demonstrate the utility of our approach, we evaluate it on different ML models in the context of image recognition. The empirical results indicate that our proposed approach not only outperforms the state-of-the-art techniques in explaining individual decisions but also provides users with an ability to discover the vulnerabilities of the target ML models.

Data-Driven Clustering via Parameterized Lloyd's Families

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #160

Maria-Florina Balcan · Travis Dick · Colin White

Algorithms for clustering points in metric spaces is a long-studied area of research. Clustering has seen a multitude of work both theoretically, in understanding the approximation guarantees possible for many objective functions such as k-median and k-means clustering, and experimentally, in finding the fastest algorithms and seeding procedures for Lloyd's algorithm. The performance of a given clustering algorithm depends on the specific application at hand, and this may not be known up front. For example, a "typical instance" may vary depending on the application, and different clustering heuristics perform differently depending on the instance. In this paper, we define an infinite family of algorithms generalizing Lloyd's algorithm, with one parameter controlling the the initialization procedure, and another parameter controlling the local search procedure. This family of algorithms includes the celebrated k-means++ algorithm, as well as the classic farthest-first traversal algorithm. We design efficient learning algorithms which receive samples from an application-specific distribution over clustering instances and learn a near-optimal clustering algorithm from the class. We show the best parameters vary significantly across datasets such as MNIST, CIFAR, and mixtures of Gaussians. Our learned algorithms never perform worse than k-means++, and on some datasets we see significant improvements.

Manifold Structured Prediction

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #161

Alessandro Rudi · Carlo Ciliberto · GianMaria Marconi · Lorenzo Rosasco

Structured prediction provides a general framework to deal with supervised problems where the outputs have semantically rich structure. While classical approaches consider finite, albeit potentially huge, output spaces, in this paper we discuss how structured prediction can be extended to a continuous scenario. Specifically, we study a structured prediction approach to manifold-valued regression. We characterize a class of problems for which the considered approach is statistically consistent and study how geometric optimization can be used to compute the corresponding estimator. Promising experimental results on both simulated and real data complete our study.

Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #162

Timur Garipov · Pavel Izmailov · Dmitrii Podoprikhin · Dmitry Vetrov · Andrew Wilson

The loss functions of deep neural networks are complex and their geometric properties are not well understood. We show that the optima of these complex loss functions are in fact connected by simple curves, over which training and test accuracy are nearly constant. We introduce a training procedure to discover these high-accuracy pathways between modes. Inspired by this new geometric insight, we also propose a new ensembling method entitled Fast Geometric Ensembling (FGE). Using FGE we can train high-performing ensembles in the time required to train a single model. We achieve improved performance compared to the recent state-of-the-art Snapshot Ensembles, on CIFAR-10, CIFAR-100, and ImageNet.

Masking: A New Perspective of Noisy Supervision

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #163

Bo Han · Jiangchao Yao · Gang Niu · Mingyuan Zhou · Ivor Tsang · Ya Zhang · Masashi Sugiyama

It is important to learn various types of classifiers given training data with noisy labels. Noisy labels, in the most popular noise model hitherto, are corrupted from ground-truth labels by an unknown noise transition matrix. Thus, by estimating this matrix, classifiers can escape from overfitting those noisy labels. However, such estimation is practically difficult, due to either the indirect nature of two-step approaches, or not big enough data to afford end-to-end approaches. In this paper, we propose a human-assisted approach called ''Masking'' that conveys human cognition of invalid class transitions and naturally speculates the structure of the noise transition matrix. To this end, we derive a structure-aware probabilistic model incorporating a structure prior, and solve the challenges from structure extraction and structure alignment. Thanks to Masking, we only estimate unmasked noise transition probabilities and the burden of estimation is tremendously reduced. We conduct extensive experiments on CIFAR-10 and CIFAR-100 with three noise structures as well as the industrial-level Clothing1M with agnostic noise structure, and the results show that Masking can improve the robustness of classifiers significantly.

Supervising Unsupervised Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #164

Vikas Garg · Adam Kalai

We introduce a framework to transfer knowledge acquired from a repository of (heterogeneous) supervised datasets to new unsupervised datasets. Our perspective avoids the subjectivity inherent in unsupervised learning by reducing it to supervised learning, and provides a principled way to evaluate unsupervised algorithms. We demonstrate the versatility of our framework via rigorous agnostic bounds on a variety of unsupervised problems. In the context of clustering, our approach helps choose the number of clusters and the clustering algorithm, remove the outliers, and provably circumvent Kleinberg's impossibility result. Experiments across hundreds of problems demonstrate improvements in performance on unsupervised data with simple algorithms despite the fact our problems come from heterogeneous domains. Additionally, our framework lets us leverage deep networks to learn common features across many small datasets, and perform zero shot learning.

One-Shot Unsupervised Cross Domain Translation

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #165

Sagie Benaim · Lior Wolf

Given a single image \$x\$ from domain \$A\$ and a set of images from domain \$B\$, our task is to generate the analogous of \$x\$ in \$B\$. We argue that this task could be a key AI capability that underlines the ability of cognitive agents to act in the world and present empirical evidence that the existing unsupervised domain translation methods fail on this task. Our method follows a two step process. First, a variational autoencoder for domain \$B\$ is trained. Then, given the new sample \$x\$, we create a variational autoencoder for domain \$A\$ by adapting the layers that are close to the image in order to directly fit \$x\$, and only indirectly adapt the other layers. Our experiments indicate that the new method does as well, when trained on one sample \$x\$, as the existing domain transfer methods, when these enjoy a multitude of training samples from domain \$A\$. Our code is made publicly available at https://github.com/sagiebenaim/OneShotTranslation

Neural Architecture Search with Bayesian Optimisation and Optimal Transport

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #166

Kirthevasan Kandasamy · Willie Neiswanger · Jeff Schneider · Barnabas Poczos · Eric Xing

Bayesian Optimisation (BO) refers to a class of methods for global optimisation of a function f which is only accessible via point evaluations. It is typically used in settings where f is expensive to evaluate. A common use case for BO in machine learning is model selection, where it is not possible to analytically model the generalisation performance of a statistical model, and we resort to noisy and expensive training and validation procedures to choose the best model. Conventional BO methods have focused on Euclidean and categorical domains, which, in the context of model selection, only permits tuning scalar hyper-parameters of machine learning algorithms. However, with the surge of interest in deep learning, there is an increasing demand to tune neural network architectures. In this work, we develop NASBOT, a Gaussian process based BO framework for neural architecture search. To accomplish this, we develop a distance metric in the space of neural network architectures which can be computed efficiently via an optimal transport program. This distance might be of independent interest to the deep learning community as it may find applications outside of BO. We demonstrate that NASBOT outperforms other alternatives for architecture search in several cross validation based model selection tasks on multi-layer perceptrons and convolutional neural networks.

Adapted Deep Embeddings: A Synthesis of Methods for k-Shot Inductive Transfer Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #167

Tyler Scott · Karl Ridgeway · Michael Mozer

The focus in machine learning has branched beyond training classifiers on a single task to investigating how previously acquired knowledge in a source domain can be leveraged to facilitate learning in a related target domain, known as inductive transfer learning. Three active lines of research have independently explored transfer learning using neural networks. In weight transfer, a model trained on the source domain is used as an initialization point for a network to be trained on the target domain. In deep metric learning, the source domain is used to construct an embedding that captures class structure in both the source and target domains. In few-shot learning, the focus is on generalizing well in the target domain based on a limited number of labeled examples. We compare state-of-the-art methods from these three paradigms and also explore hybrid adapted-embedding methods that use limited target-domain data to fine tune embeddings constructed from source-domain data. We conduct a systematic comparison of methods in a variety of domains, varying the number of labeled instances available in the target domain (k), as well as the number of target-domain classes. We reach three principal conclusions: (1) Deep embeddings are far superior, compared to weight transfer, as a starting point for inter-domain transfer or model re-use (2) Our hybrid methods robustly outperform every few-shot learning and every deep metric learning method previously proposed, with a mean error reduction of 34% over state-of-the-art. (3) Among loss functions for discovering embeddings, the histogram loss (Ustinova & Lempitsky, 2016) is most robust. We hope our results will motivate a unification of research in weight transfer, deep metric learning, and few-shot learning.

Completing State Representations using Spectral Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #168

Nan Jiang · Alex Kulesza · Satinder Singh

A central problem in dynamical system modeling is state discovery—that is, finding a compact summary of the past that captures the information needed to predict the future. Predictive State Representations (PSRs) enable clever spectral methods for state discovery; however, while consistent in the limit of infinite data, these methods often suffer from poor performance in the low data regime. In this paper we develop a novel algorithm for incorporating domain knowledge, in the form of an imperfect state representation, as side information to speed spectral learning for PSRs. We prove theoretical results characterizing the relevance of a user-provided state representation, and design spectral algorithms that can take advantage of a relevant representation. Our algorithm utilizes principal angles to extract the relevant components of the representation, and is robust to misspecification. Empirical evaluation on synthetic HMMs, an aircraft identification domain, and a gene splice dataset shows that, even with weak domain knowledge, the algorithm can significantly outperform standard PSR learning.

Data-Efficient Hierarchical Reinforcement Learning

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #169

Ofir Nachum · Shixiang (Shane) Gu · Honglak Lee · Sergey Levine

Hierarchical reinforcement learning (HRL) is a promising approach to extend traditional reinforcement learning (RL) methods to solve more complex tasks. Yet, the majority of current HRL methods require careful task-specific design and on-policy training, making them difficult to apply in real-world scenarios. In this paper, we study how we can develop HRL algorithms that are general, in that they do not make onerous additional assumptions beyond standard RL algorithms, and efficient, in the sense that they can be used with modest numbers of interaction samples, making them suitable for real-world problems such as robotic control. For generality, we develop a scheme where lower-level controllers are supervised with goals that are learned and proposed automatically by the higher-level controllers. To address efficiency, we propose to use off-policy experience for both higher- and lower-level training. This poses a considerable challenge, since changes to the lower-level behaviors change the action space for the higher-level policy, and we introduce an off-policy correction to remedy this challenge. This allows us to take advantage of recent advances in off-policy model-free RL to learn both higher and lower-level policies using substantially fewer environment interactions than on-policy algorithms. We find that our resulting HRL agent is generally applicable and highly sample-efficient. Our experiments show that our method can be used to learn highly complex behaviors for simulated robots, such as pushing objects and utilizing them to reach target locations, learning from only a few million samples, equivalent to a few days of real-time interaction. In comparisons with a number of prior HRL methods, we find that our approach substantially outperforms previous state-of-the-art techniques.

The Cluster Description Problem - Complexity Results, Formulations and Approximations

Poster | Tue Dec 4th 05:00 -- 07:00 PM @ Room 210 & 230 AB #170

Ian Davidson · Antoine Gourru · S Ravi

Consider the situation where you are given an existing k-way clustering π. A challenge for explainable AI is to find a compact and distinct explanations of each cluster which in this paper is using instance-level descriptors/tags from a common dictionary. Since the descriptors/tags were not given to the clustering method, this is not a semi-supervised learning situation. We show that the \emph{feasibility} problem of just testing whether any distinct description (not the most compact) exists is generally intractable for just two clusters. This means that unless \textbf{P} = \cnp, there cannot exist an efficient algorithm for the cluster description problem. Hence, we explore ILP formulations for smaller problems and a relaxed but restricted setting that leads to a polynomial time algorithm for larger problems. We explore several extension to the basic setting such as the ability to ignore some instances and composition constraints on the descriptions of the clusters. We show our formulation's usefulness on Twitter data where the communities were found using social connectivity (i.e. \texttt{follower} relation) but the explanation of the communities is based on behavioral properties of the nodes (i.e. hashtag usage) not available to the clustering method.