

Synthesized Policies for Transfer and Adaptation across Tasks and Environments

Hexiang Hu, Liyu Chen, Boqing Gong, Fei Sha

The ability to transfer in reinforcement learning is key towards building an agent of general artificial intelligence. In this paper, we consider the problem of learning to simultaneously transfer across both environments and tasks, probably more importantly, by learning from only sparse (environment, task) pairs out of all the possible combinations. We propose a novel compositional neural network architecture which depicts a meta rule for composing policies from environment and task embeddings. Notably, one of the main challenges is to learn the embeddings jointly with the meta rule. We further propose new training methods to disentangle the embeddings, making them both distinctive signatures of the environments and tasks and effective building blocks for composing the policies. Experiments on GridWorld and THOR, of which the agent takes as input an egocentric view, show that our approach gives rise to high success rates on all the (environment, task) pairs after learning from only 40% of them.

Self-Supervised Generation of Spatial Audio for 360° Video

Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, Oliver Wang

We introduce an approach to convert mono audio recorded by a 360° video camera into spatial audio, a representation of the distribution of sound over the full viewing sphere. Spatial audio is an important component of immersive 360° video viewing, but spatial audio microphones are still rare in current 360° video production. Our system consists of end-to-end trainable neural networks that separate individual sound sources and localize them on the viewing sphere, conditioned on multi-modal analysis from the audio and 360° video frames. We introduce several datasets, including one filmed ourselves, and one collected in-the-wild from YouTube, consisting of 360° videos uploaded with spatial audio. During training, ground truth spatial audio serves as self-supervision and a mixed down mono track forms the input to our network. Using our approach we show that it is possible to infer the spatial localization of sounds based only on a synchronized 360° video and the mono audio track.

On GANs and GMMs

Eitan Richardson, Yair Weiss

A longstanding problem in machine learning is to find unsupervised methods that can learn the statistical structure of high dimensional signals. In recent years, GANs have gained much attention as a possible solution to the problem, and in particular have shown the ability to generate remarkably realistic high resolution sampled images. At the same time, many authors have pointed out that GANs may fail to model the full distribution ("mode collapse") and that using the learned models for anything other than generating samples may be very difficult.

Batch-Instance Normalization for Adaptively Style-Invariant Neural Networks

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 trained 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.

Hierarchical Reinforcement Learning for Zero-shot Generalization with Subtask Dependencies

SungrYull Sohn, Junhyuk Oh, Honglak Lee

We introduce a new RL problem where the agent is required to generalize to a previously-unseen environment characterized by a subtask graph which describes a set of subtasks and their dependencies. Unlike existing hierarchical multitask RL approaches that explicitly describe what the agent should do at a high level, our problem only describes properties of subtasks and relationships among them, which requires the agent to perform complex reasoning to find the optimal subtask to execute. To solve this problem, we propose a neural subtask graph solver (NSGS) which encodes the subtask graph using a recursive neural network embedding. To overcome the difficulty of training, we propose a novel non-parametric gradient-based policy, graph reward propagation, to pre-train our NSGS agent and further finetune it through actor-critic method. The experimental results on two 2D visual domains show that our agent can perform complex reasoning to find a near-optimal way of executing the subtask graph and generalize well to the unseen subtask graphs. In addition, we compare our agent with a Monte-Carlo tree search (MCTS) method showing that our method is much more efficient than MCTS, and the performance of NSGS can be further improved by combining it with MCTS.

KDGAN: Knowledge Distillation with Generative Adversarial Networks

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.

Contour location via entropy reduction leveraging multiple information sources

Alexandre Marques, Remi Lam, Karen Willcox

We introduce an algorithm to locate contours of functions that are expensive to evaluate. The problem of locating contours arises in many applications, including classification, constrained optimization, and performance analysis of mechanical and dynamical systems (reliability, probability of failure, stability, etc.). Our algorithm locates contours using information from multiple sources, which are available in the form of relatively inexpensive, biased, and possibly noisy approximations to the original function. Considering multiple information sources can lead to significant cost savings. We also introduce the concept of contour entropy, a formal measure of uncertainty about the location of the zero contour of a function approximated by a statistical surrogate model. Our algorithm locates contours efficiently by maximizing the reduction of contour entropy per unit cost.

Contextual bandits with surrogate losses: Margin bounds and efficient algorithms
Dylan J. Foster, Akshay Krishnamurthy

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Adaptive Sampling Towards Fast Graph Representation Learning

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.

Analytic solution and stationary phase approximation for the Bayesian lasso and elastic net

Tom Michoel

The lasso and elastic net linear regression models impose a double-exponential prior distribution on the model parameters to achieve regression shrinkage and variable selection, allowing the inference of robust models from large data sets. However, there has been limited success in deriving estimates for the full posterior distribution of regression coefficients in these models, due to a need to evaluate analytically intractable partition function integrals. Here, the Fourier transform is used to express these integrals as complex-valued oscillatory integrals over "regression frequencies". This results in an analytic expansion and stationary phase approximation for the partition functions of the Bayesian lasso and elastic net, where the non-differentiability of the double-exponential prior has so far eluded such an approach. Use of this approximation leads to highly accurate numerical estimates for the expectation values and marginal posterior distributions of the regression coefficients, and allows for Bayesian inference of much higher dimensional models than previously possible.

Identification and Estimation of Causal Effects from Dependent Data

Eli Sherman, Ilya Shpitser

The assumption that data samples are independent and identically distributed (iid) is standard in many areas of statistics and machine learning. Nevertheless, in some settings, such as social networks, infectious disease modeling, and reasoning with spatial and temporal data, this assumption is false. An extensive literature exists on making causal inferences under the iid assumption [12, 8, 21, 16], but, as pointed out in [14], causal inference in non-iid contexts is challenging due to the combination of unobserved confounding bias and data dependence. In this paper we develop a general theory describing when causal inferences are possible in such scenarios. We use segregated graphs [15], a generalization of latent projection mixed graphs [23], to represent causal models of this type and provide a complete algorithm for non-parametric identification in these models. We then demonstrate how statistical inferences may be performed on causal parameters identified by this algorithm, even in cases where parts of the model exhibit full interference, meaning only a single sample is available for parts of the

model [19]. We apply these techniques to a synthetic data set which considers the adoption of fake news articles given the social network structure, articles read by each person, and baseline demographics and socioeconomic covariates.

Streamlining Variational Inference for Constraint Satisfaction Problems

Aditya Grover, Tudor Achim, Stefano Ermon

Several algorithms for solving constraint satisfaction problems are based on survey propagation, a variational inference scheme used to obtain approximate marginal probability estimates for variable assignments. These marginals correspond to how frequently each variable is set to true among satisfying assignments, and are used to inform branching decisions during search; however, marginal estimates obtained via survey propagation are approximate and can be self-contradictory.

We introduce a more general branching strategy based on streamlining constraints, which sidestep hard assignments to variables. We show that streamlined solvers consistently outperform decimation-based solvers on random k-SAT instances for several problem sizes, shrinking the gap between empirical performance and theoretical limits of satisfiability by 16.3% on average for $k = 3, 4, 5, 6$.

A Spectral View of Adversarially Robust Features

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.

Dimensionality Reduction has Quantifiable Imperfections: Two Geometric Bounds

Kry Lui, Gavin Weiguang Ding, Ruitong Huang, Robert McCann

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning SMaLL Predictors

Vikas Garg, Ofer Dekel, Lin Xiao

We introduce a new framework for learning in severely resource-constrained settings. Our technique delicately amalgamates the representational richness of multiple linear predictors with the sparsity of Boolean relaxations, and thereby yields classifiers that are compact, interpretable, and accurate. We provide a rigorous formalism of the learning problem, and establish fast convergence of the ensuing algorithm via relaxation to a minimax saddle point objective. We supplement the theoretical foundations of our work with an extensive empirical evaluation.

ResNet with one-neuron hidden layers is a Universal Approximator

Hongzhou Lin, Stefanie Jegelka

We demonstrate that a very deep ResNet with stacked modules that have one neuron per hidden layer and ReLU activation functions can uniformly approximate any Lebesgue integrable function in d dimensions, i.e. $\ell_1(R^d)$. Due to the identity mapping inherent to ResNets, our network has alternating layers of dimension one and d . This stands in sharp contrast to fully connected networks, which are not universal approximators if their width is the input dimension d [21,11]. Hence, our result implies an increase in representational power for narrow deep networks by the ResNet architecture.

How Many Samples are Needed to Estimate a Convolutional Neural Network?

Simon S. Du, Yining Wang, Xiyu Zhai, Sivaraman Balakrishnan, Russ R. Salakhutdinov, Aarti Singh

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Objective and efficient inference for couplings in neuronal networks

Yu Terada, Tomoyuki Obuchi, Takuya Isomura, Yoshiyuki Kabashima

Inferring directional couplings from the spike data of networks is desired in various scientific fields such as neuroscience. Here, we apply a recently proposed objective procedure to the spike data obtained from the Hodgkin-Huxley type models and in vitro neuronal networks cultured in a circular structure. As a result, we succeed in reconstructing synaptic connections accurately from the evoked activity as well as the spontaneous one. To obtain the results, we invent an analytic formula approximately implementing a method of screening relevant couplings. This significantly reduces the computational cost of the screening method employed in the proposed objective procedure, making it possible to treat large-size systems as in this study.

Unsupervised Adversarial Invariance

Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, Prem Natarajan

Data representations that contain all the information about target variables but are invariant to nuisance factors benefit supervised learning algorithms by preventing them from learning associations between these factors and the targets, thus reducing overfitting. We present a novel unsupervised invariance induction framework for neural networks that learns a split representation of data through competitive training between the prediction task and a reconstruction task coupled with disentanglement, without needing any labeled information about nuisance factors or domain knowledge. We describe an adversarial instantiation of this framework and provide analysis of its working. Our unsupervised model outperforms state-of-the-art methods, which are supervised, at inducing invariance to inherent nuisance factors, effectively using synthetic data augmentation to learn invariance, and domain adaptation. Our method can be applied to any prediction task, eg., binary/multi-class classification or regression, without loss of generality.

Critical initialisation for deep signal propagation in noisy rectifier neural networks

Arnu Pretorius, Elan van Biljon, Steve Kroon, Herman Kamper

Stochastic regularisation is an important weapon in the arsenal of a deep learning practitioner. However, despite recent theoretical advances, our understanding of how noise influences signal propagation in deep neural networks remains limited. By extending recent work based on mean field theory, we develop a new framework for signal propagation in stochastic regularised neural networks. Our \textit{noisy signal propagation} theory can incorporate several common noise distributions, including additive and multiplicative Gaussian noise as well as dropout.

We use this framework to investigate initialisation strategies for noisy ReLU networks. We show that no critical initialisation strategy exists using additive noise, with signal propagation exploding regardless of the selected noise distribution. For multiplicative noise (e.g. dropout), we identify alternative critical initialisation strategies that depend on the second moment of the noise distribution. Simulations and experiments on real-world data confirm that our proposed initialisation is able to stably propagate signals in deep networks, while using an initialisation disregarding noise fails to do so. Furthermore, we analyse correlation dynamics between inputs. Stronger noise regularisation is shown to reduce the depth to which discriminatory information about the inputs to a noisy ReLU network is able to propagate, even when initialised at criticality. We sup

port our theoretical predictions for these trainable depths with simulations, as well as with experiments on MNIST and CIFAR-10.

Learning sparse neural networks via sensitivity-driven regularization

Enzo Tartaglione, Skjalg Lepsøy, Attilio Fiandrotti, Gianluca Francini

The ever-increasing number of parameters in deep neural networks poses challenges for memory-limited applications. Regularize-and-prune methods aim at meeting these challenges by sparsifying the network weights. In this context we quantify the output sensitivity to the parameters (i.e. their relevance to the network output) and introduce a regularization term that gradually lowers the absolute value of parameters with low sensitivity. Thus, a very large fraction of the parameters approach zero and are eventually set to zero by simple thresholding. Our method surpasses most of the recent techniques both in terms of sparsity and error rates. In some cases, the method reaches twice the sparsity obtained by other techniques at equal error rates.

Cooperative neural networks (CoNN): Exploiting prior independence structure for improved classification

Harsh Shrivastava, Eugene Bart, Bob Price, Hanjun Dai, Bo Dai, Srinivas Aluru

We propose a new approach, called cooperative neural networks (CoNN), which use a set of cooperatively trained neural networks to capture latent representations that exploit prior given independence structure. The model is more flexible than traditional graphical models based on exponential family distributions, but incorporates more domain specific prior structure than traditional deep networks or variational autoencoders. The framework is very general and can be used to exploit the independence structure of any graphical model. We illustrate the technique by showing that we can transfer the independence structure of the popular Latent Dirichlet Allocation (LDA) model to a cooperative neural network, CoNN-sLDA. Empirical evaluation of CoNN-sLDA on supervised text classification tasks demonstrate that the theoretical advantages of prior independence structure can be realized in practice - we demonstrate a 23 percent reduction in error on the challenging MultiSent data set compared to state-of-the-art.

Data center cooling using model-predictive control

Nevena Lazic, Craig Boutilier, Tyler Lu, Eehern Wong, Binz Roy, MK Ryu, Greg Imwalleye

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 air flow 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.

Generalization Bounds for Uniformly Stable Algorithms

Vitaly Feldman, Jan Vondrak

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

COLA: Decentralized Linear Learning

Lie He, An Bian, Martin Jaggi

Decentralized machine learning is a promising emerging paradigm in view of global challenges of data ownership and privacy. We consider learning of linear classification and regression models, in the setting where the training data is decentralized over many user devices, and the learning algorithm must run on-device, on an arbitrary communication network, without a central coordinator.

We propose COLA, a new decentralized training algorithm with strong theoretical guarantees and superior practical performance. Our framework overcomes many limitations of existing methods, and achieves communication efficiency, scalability, elasticity as well as resilience to changes in data and allows for unreliable and heterogeneous participating devices.

Leveraging the Exact Likelihood of Deep Latent Variable Models

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.

A General Method for Amortizing Variational Filtering

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.

One-Shot Unsupervised Cross Domain Translation

Sagie Benaïm, Lior Wolf

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Query K-means Clustering and the Double Dixie Cup Problem

I Chien, Chao Pan, Olgica Milenkovic

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Probabilistic Neural Programmed Networks for Scene Generation

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.

Escaping Saddle Points in Constrained Optimization

Aryan Mokhtari, Asuman Ozdaglar, Ali Jadbabaie

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Adversarial Text Generation via Feature-Mover's Distance

Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, Lawrence Carin

Generative adversarial networks (GANs) have achieved significant success in generating real-valued data. However, the discrete nature of text hinders the application of GAN to text-generation tasks. Instead of using the standard GAN objective, we propose to improve text-generation GAN via a novel approach inspired by optimal transport. Specifically, we consider matching the latent feature distributions of real and synthetic sentences using a novel metric, termed the feature-mover's distance (FMD). This formulation leads to a highly discriminative critic and easy-to-optimize objective, overcoming the mode-collapsing and brittle-training problems in existing methods. Extensive experiments are conducted on a variety of tasks to evaluate the proposed model empirically, including unconditional text generation, style transfer from non-parallel text, and unsupervised cipher cracking. The proposed model yields superior performance, demonstrating wide applicability and effectiveness.

On gradient regularizers for MMD GANs

Michael Arbel, Danica J. Sutherland, Mikołaj Bińkowski, Arthur Gretton

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Differentially Private Bayesian Inference for Exponential Families

Garrett Bernstein, Daniel R. Sheldon

The study of private inference has been sparked by growing concern regarding the analysis of data when it stems from sensitive sources. We present the first method for private Bayesian inference in exponential families that properly accounts for noise introduced by the privacy mechanism. It is efficient because it works only with sufficient statistics and not individual data. Unlike other methods, it gives properly calibrated posterior beliefs in the non-asymptotic data regime.

Stochastic Primal-Dual Method for Empirical Risk Minimization with $O(1)$ Per-Iteration Complexity

Conghui Tan, Tong Zhang, Shiqian Ma, Ji Liu

Regularized empirical risk minimization problem with linear predictor appears frequently in machine learning. In this paper, we propose a new stochastic primal-dual method to solve this class of problems. Different from existing methods, our proposed methods only require $O(1)$ operations in each iteration. We also develop a variance-reduction variant of the algorithm that converges linearly. Numerical experiments suggest that our methods are faster than existing ones such as proximal SGD, SVRG and SAGA on high-dimensional problems.

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

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.

Learning Plannable Representations with Causal InfoGAN

Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart J. Russell, Pieter Abbeel

In recent years, deep generative models have been shown to 'imagine' convincing high-dimensional observations such as images, audio, and even video, learning directly from raw data. In this work, we ask how to imagine goal-directed visual plans -- a plausible sequence of observations that transition a dynamical system from its current configuration to a desired goal state, which can later be used as a reference trajectory for control. We focus on systems with high-dimensional observations, such as images, and propose an approach that naturally combines representation learning and planning. Our framework learns a generative model of sequential observations, where the generative process is induced by a transition in a low-dimensional planning model, and an additional noise. By maximizing the mutual information between the generated observations and the transition in the planning model, we obtain a low-dimensional representation that best explains the causal nature of the data. We structure the planning model to be compatible with efficient planning algorithms, and we propose several such models based on either discrete or continuous states. Finally, to generate a visual plan, we project the current and goal observations onto their respective states in the planning model, plan a trajectory, and then use the generative model to transform the trajectory to a sequence of observations. We demonstrate our method on imagining plausible visual plans of rope manipulation.

Interactive Structure Learning with Structural Query-by-Committee

Christopher Tosh, Sanjoy Dasgupta

In this work, we introduce interactive structure learning, a framework that unifies many different interactive learning tasks. We present a generalization of the query-by-committee active learning algorithm for this setting, and we study its consistency and rate of convergence, both theoretically and empirically, with and without noise.

The streaming rollout of deep networks - towards fully model-parallel execution

Volker Fischer, Jan Koehler, Thomas Pfeil

Deep neural networks, and in particular recurrent networks, are promising candidates to control autonomous agents that interact in real-time with the physical world. However, this requires a seamless integration of temporal features into the network's architecture. For the training of and inference with recurrent neural networks, they are usually rolled out over time, and different rollouts exist.

Conventionally during inference, the layers of a network are computed in a sequential manner resulting in sparse temporal integration of information and long r

response times. In this study, we present a theoretical framework to describe rollouts, the level of model-parallelization they induce, and demonstrate differences in solving specific tasks. We prove that certain rollouts, also for networks with only skip and no recurrent connections, enable earlier and more frequent responses, and show empirically that these early responses have better performance. The streaming rollout maximizes these properties and enables a fully parallel execution of the network reducing runtime on massively parallel devices. Finally, we provide an open-source toolbox to design, train, evaluate, and interact with streaming rollouts.

Contextual Stochastic Block Models

Yash Deshpande, Subhabrata Sen, Andrea Montanari, Elchanan Mossel

We provide the first information theoretical tight analysis for inference of latent community structure given a sparse graph along with high dimensional node covariates, correlated with the same latent communities. Our work bridges recent theoretical breakthroughs in detection of latent community structure without node covariates and a large body of empirical work using diverse heuristics for combining node covariates with graphs for inference. The tightness of our analysis implies in particular, the information theoretic necessity of combining the different sources of information.

Our analysis holds for networks of large degrees as well as for a Gaussian version of the model.

Unsupervised Learning of Artistic Styles with Archetypal Style Analysis

Daan Wynen, Cordelia Schmid, Julien Mairal

In this paper, we introduce an unsupervised learning approach to automatically discover, 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.

Generalisation in humans and deep neural networks

Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, Felix A. Wichmann

We compare the robustness of humans and current convolutional deep neural networks (DNNs) on object recognition under twelve different types of image degradations. First, using three well known DNNs (ResNet-152, VGG-19, GoogLeNet) we find the human visual system to be more robust to nearly all of the tested image manipulations, and we observe progressively diverging classification error-patterns between humans and DNNs when the signal gets weaker. Secondly, we show that DNNs trained directly on distorted images consistently surpass human performance on the exact distortion types they were trained on, yet they display extremely poor generalisation abilities when tested on other distortion types. For example, training on salt-and-pepper noise does not imply robustness on uniform white noise and vice versa. Thus, changes in the noise distribution between training and testing constitutes a crucial challenge to deep learning vision systems that can be

systematically addressed in a lifelong machine learning approach. Our new dataset consisting of 83K carefully measured human psychophysical trials provide a useful reference for lifelong robustness against image degradations set by the human visual system.

Distributed Weight Consolidation: A Brain Segmentation Case Study

Patrick McClure, Charles Y. Zheng, Jakub Kaczmarzyk, John Rogers-Lee, Satra Ghosh, Dylan Nielson, Peter A. Bandettini, Francisco Pereira

Collecting the large datasets needed to train deep neural networks can be very difficult, particularly for the many applications for which sharing and pooling data is complicated by practical, ethical, or legal concerns. However, it may be the case that derivative datasets or predictive models developed within individual sites can be shared and combined with fewer restrictions. Training on distributed data and combining the resulting networks is often viewed as continual learning, but these methods require networks to be trained sequentially. In this paper, we introduce distributed weight consolidation (DWC), a continual learning method to consolidate the weights of separate neural networks, each trained on an independent dataset. We evaluated DWC with a brain segmentation case study, where we consolidated dilated convolutional neural networks trained on independent structural magnetic resonance imaging (sMRI) datasets from different sites. We found that DWC led to increased performance on test sets from the different sites, while maintaining generalization performance for a very large and completely independent multi-site dataset, compared to an ensemble baseline.

IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis

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.

MixLasso: Generalized Mixed Regression via Convex Atomic-Norm Regularization

Ian En-Hsu Yen, Wei-Cheng Lee, Kai Zhong, Sung-En Chang, Pradeep K. Ravikumar, Shou-De Lin

We consider a generalization of mixed regression where the response is an additive combination of several mixture components. Standard mixed regression is a special case where each response is generated from exactly one component. Typical approaches to the mixture regression problem employ local search methods such as Expectation Maximization (EM) that are prone to spurious local optima. On the other hand, a number of recent theoretically-motivated *Tensor-based methods* either have high sample complexity, or require the knowledge of the input distribution, which is not available in most of practical situations. In this work, we study a novel convex estimator *MixLasso* for the estimation of generalized mixed regression, based on an atomic norm specifically constructed to regularize the number of mixture components. Our algorithm gives a risk bound that trades off between prediction accuracy and model sparsity without imposing stringent assumptions on the input/output distribution, and can be easily adapted to the

case of non-linear functions. In our numerical experiments on mixtures of linear as well as nonlinear regressions, the proposed method yields high-quality solutions in a wider range of settings than existing approaches.

A Dual Framework for Low-rank Tensor Completion

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.

Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer

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.

Enhancing the Accuracy and Fairness of Human Decision Making

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?

Computing Higher Order Derivatives of Matrix and Tensor Expressions

Soeren Laue, Matthias Mitterreiter, Joachim Giesen

Optimization is an integral part of most machine learning systems and most numerical optimization schemes rely on the computation of derivatives. Therefore, frameworks for computing derivatives are an active area of machine learning research. Surprisingly, as of yet, no existing framework is capable of computing higher order matrix and tensor derivatives directly. Here, we close this fundamental gap and present an algorithmic framework for computing matrix and tensor derivatives that extends seamlessly to higher order derivatives. The framework can be used for symbolic as well as for forward and reverse mode automatic differentiation. Experiments show a speedup between one and four orders of magnitude over state-of-the-art frameworks when evaluating higher order derivatives.

Efficient online algorithms for fast-rate regret bounds under sparsity

Pierre Gaillard, Olivier Wintenberger

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Human-in-the-Loop Interpretability Prior

Isaac Lage, Andrew Ross, Samuel J. Gershman, Been Kim, Finale Doshi-Velez

We often desire our models to be interpretable as well as accurate. Prior work on optimizing models for interpretability has relied on easy-to-quantify proxies for interpretability, such as sparsity or the number of operations required. In this work, we optimize for interpretability by directly including humans in the optimization loop. We develop an algorithm that minimizes the number of user studies to find models that are both predictive and interpretable and demonstrate our approach on several data sets. Our human subjects results show trends towards different proxy notions of interpretability on different datasets, which suggests that different proxies are preferred on different tasks.

Deep Non-Blind Deconvolution via Generalized Low-Rank Approximation

Wenqi Ren, Jiawei Zhang, Lin Ma, Jinshan Pan, Xiaochun Cao, Wangmeng Zuo, Wei Liu, Ming-Hsuan Yang

In this paper, we present a deep convolutional neural network to capture the inherent properties of image degradation, which can handle different kernels and saturated pixels in a unified framework. The proposed neural network is motivated by the low-rank property of pseudo-inverse kernels. We first compute a generalized low-rank approximation for a large number of blur kernels, and then use separable filters to initialize the convolutional parameters in the network. Our analysis shows that the estimated decomposed matrices contain the most essential information of the input kernel, which ensures the proposed network to handle various blurs in a unified framework and generate high-quality deblurring results. Experimental results on benchmark datasets with noise and saturated pixels demonstrate that the proposed algorithm performs favorably against state-of-the-art methods.

Regret Bounds for Robust Adaptive Control of the Linear Quadratic Regulator

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, Stephen Tu

We consider adaptive control of the Linear Quadratic Regulator (LQR), where an unknown linear system is controlled subject to quadratic costs. Leveraging recent

developments in the estimation of linear systems and in robust controller synthesis,

we present the first provably polynomial time algorithm that achieves sub-linear regret on this problem. We further study the interplay between regret minimization

and parameter estimation by proving a lower bound on the expected regret in terms of the exploration schedule used by any algorithm. Finally, we conduct a numerical study comparing our robust adaptive algorithm to other methods from the adaptive LQR literature, and demonstrate the flexibility of our proposed method

by extending it to a demand forecasting problem subject to state constraints.

Binary Rating Estimation with Graph Side Information

Kwangjun Ahn, Kangwook Lee, Hyunseung Cha, Changho Suh

Rich experimental evidences show that one can better estimate users' unknown ratings with the aid of graph side information such as social graphs. However, the gain is not theoretically quantified. In this work, we study the binary rating estimation problem to understand the fundamental value of graph side information.

Considering a simple correlation model between a rating matrix and a graph, we characterize the sharp threshold on the number of observed entries required to recover the rating matrix (called the optimal sample complexity) as a function of

the quality of graph side information (to be detailed). To the best of our knowledge, we are the first to reveal how much the graph side information reduces sample complexity. Further, we propose a computationally efficient algorithm that achieves the limit. Our experimental results demonstrate that the algorithm performs well even with real-world graphs.

A Bayesian Nonparametric View on Count-Min Sketch

Diana Cai, Michael Mitzenmacher, Ryan P. Adams

The count-min sketch is a time- and memory-efficient randomized data structure that provides a point estimate of the number of times an item has appeared in a data stream. The count-min sketch and related hash-based data structures are ubiquitous in systems that must track frequencies of data such as URLs, IP addresses, and language n-grams. We present a Bayesian view on the count-min sketch, using the same data structure, but providing a posterior distribution over the frequencies that characterizes the uncertainty arising from the hash-based approximation. In particular, we take a nonparametric approach and consider tokens generated from a Dirichlet process (DP) random measure, which allows for an unbounded number of unique tokens. Using properties of the DP, we show that it is possible to straightforwardly compute posterior marginals of the unknown true counts and that the modes of these marginals recover the count-min sketch estimator, inheriting the associated probabilistic guarantees. Using simulated data with known ground truth, we investigate the properties of these estimators. Lastly, we also study a modified problem in which the observation stream consists of collections of tokens (i.e., documents) arising from a random measure drawn from a stable beta process, which allows for power law scaling behavior in the number of unique tokens.

Provable Variational Inference for Constrained Log-Submodular Models

Josip Djolonga, Stefanie Jegelka, Andreas Krause

Submodular maximization problems appear in several areas of machine learning and data science, as many useful modelling concepts such as diversity and coverage satisfy this natural diminishing returns property. Because the data defining these functions, as well as the decisions made with the computed solutions, are subject to statistical noise and randomness, it is arguably necessary to go beyond computing a single approximate optimum and quantify its inherent uncertainty. To this end, we define a rich class of probabilistic models associated with constrained submodular maximization problems. These capture log-submodular dependencies of arbitrary order between the variables, but also satisfy hard combinatorial constraints. Namely, the variables are assumed to take on one of m – possibly exponentially many – set of states, which form the bases of a matroid. To perform inference in these models we design novel variational inference algorithms, which carefully leverage the combinatorial and probabilistic properties of these objects. In addition to providing completely tractable and well-understood variational approximations, our approach results in the minimization of a convex upper bound on the log-partition function. The bound can be efficiently evaluated using greedy algorithms and optimized using any first-order method. Moreover, for the case of facility location and weighted coverage functions, we prove the first constant factor guarantee in this setting – an efficiently certifiable $e/(e-1)$ approximation of the log-partition function. Finally, we empirically demonstrate the effectiveness of our approach on several instances.

Middle-Out Decoding

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.

ons. 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.

Maximum-Entropy Fine Grained Classification

Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, Nikhil Naik

Fine-Grained Visual Classification (FGVC) is an important computer vision problem that involves small diversity within the different classes, and often requires expert annotators to collect data. Utilizing this notion of small visual diversity, we revisit Maximum-Entropy learning in the context of fine-grained classification, and provide a training routine that maximizes the entropy of the output probability distribution for training convolutional neural networks on FGVC tasks. We provide a theoretical as well as empirical justification of our approach, and achieve state-of-the-art performance across a variety of classification tasks in FGVC, that can potentially be extended to any fine-tuning task. Our method is robust to different hyperparameter values, amount of training data and amount of training label noise and can hence be a valuable tool in many similar problems.

Deep State Space Models for Unconditional Word Generation

Florian Schmidt, Thomas Hofmann

Autoregressive feedback is considered a necessity for successful unconditional text generation using stochastic sequence models. However, such feedback is known to introduce systematic biases into the training process and it obscures a principle of generation: committing to global information and forgetting local nuances. We show that a non-autoregressive deep state space model with a clear separation of global and local uncertainty can be built from only two ingredients: An independent noise source and a deterministic transition function. Recent advances on flow-based variational inference can be used to train an evidence lower-bound without resorting to annealing, auxiliary losses or similar measures. The result is a highly interpretable generative model on par with comparable autoregressive models on the task of word generation.

Total stochastic gradient algorithms and applications in reinforcement learning

Paavo Parmas

Backpropagation and the chain rule of derivatives have been prominent; however, the total derivative rule has not enjoyed the same amount of attention. In this work

we show how the total derivative rule leads to an intuitive visual framework for creating gradient estimators on graphical models. In particular, previous "policy

gradient theorems" are easily derived. We derive new gradient estimators based on density estimation, as well as a likelihood ratio gradient, which "jumps" to an

intermediate node, not directly to the objective function. We evaluate our methods

on model-based policy gradient algorithms, achieve good performance, and present evidence towards demystifying the success of the popular PILCO algorithm.

Neural Arithmetic Logic Units

Andrew Trask, Felix Hill, Scott E. Reed, Jack Rae, Chris Dyer, Phil Blunsom

Neural networks can learn to represent and manipulate numerical information, but they seldom generalize well outside of the range of numerical values encountered during training. To encourage more systematic numerical extrapolation, we propose an architecture that represents numerical quantities as linear activations w

hich are manipulated using primitive arithmetic operators, controlled by learned gates. We call this module a neural arithmetic logic unit (NALU), by analogy to the arithmetic logic unit in traditional processors. Experiments show that NALU-enhanced neural networks can learn to track time, perform arithmetic over images of numbers, translate numerical language into real-valued scalars, execute computer code, and count objects in images. In contrast to conventional architectures, we obtain substantially better generalization both inside and outside of the range of numerical values encountered during training, often extrapolating orders of magnitude beyond trained numerical ranges.

Implicit Bias of Gradient Descent on Linear Convolutional Networks

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, Nati Srebro

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

On Binary Classification in Extreme Regions

Hamid JALALZAI, Stephan Cl  men  on, Anne Sabourin

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Adaptive Skip Intervals: Temporal Abstraction for Recurrent Dynamical Models

Alexander Neitz, Giambattista Parascandolo, Stefan Bauer, Bernhard Sch  lkopf

We introduce a method which enables a recurrent dynamics model to be temporally abstract. Our approach, which we call Adaptive Skip Intervals (ASI), is based on the observation that in many sequential prediction tasks, the exact time at which events occur is irrelevant to the underlying objective. Moreover, in many situations, there exist prediction intervals which result in particularly easy-to-predict transitions. We show that there are prediction tasks for which we gain both computational efficiency and prediction accuracy by allowing the model to make predictions at a sampling rate which it can choose itself.

Learning to Specialize with Knowledge Distillation for Visual Question Answering

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 attractive 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.

A Model for Learned Bloom Filters and Optimizing by Sandwiching

Michael Mitzenmacher

Recent work has suggested enhancing Bloom filters by using a pre-filter, based on applying machine learning to determine a function that models the data set the Bloom filter is meant to represent. Here we model such learned Bloom filters,

with the following outcomes: (1) we clarify what guarantees can and cannot be associated with such a structure; (2) we show how to estimate what size the learning function must obtain in order to obtain improved performance; (3) we provide a simple method, sandwiching, for optimizing learned Bloom filters; and (4) we propose a design and analysis approach for a learned Bloomier filter, based on our modeling approach.

Random Feature Stein Discrepancies

Jonathan Huggins, Lester Mackey

Computable Stein discrepancies have been deployed for a variety of applications, ranging from sampler selection in posterior inference to approximate Bayesian inference to goodness-of-fit testing. Existing convergence-determining Stein discrepancies admit strong theoretical guarantees but suffer from a computational cost that grows quadratically in the sample size. While linear-time Stein discrepancies have been proposed for goodness-of-fit testing, they exhibit avoidable degradations in testing power—even when power is explicitly optimized. To address these shortcomings, we introduce feature Stein discrepancies (Φ SDs), a new family of quality measures that can be cheaply approximated using importance sampling. We show how to construct Φ SDs that provably determine the convergence of a sampler to its target and develop high-accuracy approximations—random Φ SDs ($R\Phi$ SDs)—which are computable in near-linear time. In our experiments with sampler selection for approximate posterior inference and goodness-of-fit testing, $R\Phi$ SDs perform as well or better than quadratic-time KSDs while being orders of magnitude faster to compute.

Nonparametric Bayesian Lomax delegate racing for survival analysis with competing risks

Quan Zhang, Mingyuan Zhou

We propose Lomax delegate racing (LDR) to explicitly model the mechanism of survival under competing risks and to interpret how the covariates accelerate or decelerate the time to event. LDR explains non-monotonic covariate effects by racing a potentially infinite number of sub-risks, and consequently relaxes the ubiquitous proportional-hazards assumption which may be too restrictive. Moreover, LDR is naturally able to model not only censoring, but also missing event times or event types. For inference, we develop a Gibbs sampler under data augmentation for moderately sized data, along with a stochastic gradient descent maximum a posteriori inference algorithm for big data applications. Illustrative experiments are provided on both synthetic and real datasets, and comparison with various benchmark algorithms for survival analysis with competing risks demonstrates distinguished performance of LDR.

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

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.

Adaptive Online Learning in Dynamic Environments

Lijun Zhang, Shiyin Lu, Zhi-Hua Zhou

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning in Games with Lossy Feedback

Zhengyuan Zhou, Panayotis Mertikopoulos, Susan Athey, Nicholas Bambos, Peter W. Glynn, Yinyu Ye

We consider a game-theoretical multi-agent learning problem where the feedback information can be lost during the learning process and rewards are given by a broad class of games known as variationally stable games. We propose a simple variant of the classical online gradient descent algorithm, called reweighted online gradient descent (ROGD) and show that in variationally stable games, if each agent adopts ROGD, then almost sure convergence to the set of Nash equilibria is guaranteed, even when the feedback loss is asynchronous and arbitrarily correlated among agents. We then extend the framework to deal with unknown feedback loss probabilities by using an estimator (constructed from past data) in its replacement. Finally, we further extend the framework to accommodate both asynchronous loss and stochastic rewards and establish that multi-agent ROGD learning still converges to the set of Nash equilibria in such settings. Together, these results contribute to the broad landscape of multi-agent online learning by significantly relaxing the feedback information that is required to achieve desirable outcomes.

Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes

Loucas Pillaud-Vivien, Alessandro Rudi, Francis Bach

We consider stochastic gradient descent (SGD) for least-squares regression with potentially several passes over the data. While several passes have been widely reported to perform practically better in terms of predictive performance on unseen data, the existing theoretical analysis of SGD suggests that a single pass is statistically optimal. While this is true for low-dimensional easy problems, we show that for hard problems, multiple passes lead to statistically optimal predictions while single pass does not; we also show that in these hard models, the optimal number of passes over the data increases with sample size. In order to define the notion of hardness and show that our predictive performances are optimal, we consider potentially infinite-dimensional models and notions typically associated to kernel methods, namely, the decay of eigenvalues of the covariance matrix of the features and the complexity of the optimal predictor as measured through the covariance matrix.

We illustrate our results on synthetic experiments with non-linear kernel methods and on a classical benchmark with a linear model.

Multimodal Generative Models for Scalable Weakly-Supervised Learning

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-supervise

d 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.

Multi-Class Learning: From Theory to Algorithm

Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, Weiping Wang

In this paper, we study the generalization performance of multi-class classification and obtain a sharper data-dependent generalization error bound with fast convergence rate, substantially improving the state-of-art bounds in the existing data-dependent generalization analysis. The theoretical analysis motivates us to devise two effective multi-class kernel learning algorithms with statistical guarantees. Experimental results show that our proposed methods can significantly outperform the existing multi-class classification methods.

Learning and Inference in Hilbert Space with Quantum Graphical Models

Siddarth Srinivasan, Carlton Downey, Byron Boots

Quantum Graphical Models (QGMs) generalize classical graphical models by adopting the formalism for reasoning about uncertainty from quantum mechanics. Unlike classical graphical models, QGMs represent uncertainty with density matrices in complex Hilbert spaces. Hilbert space embeddings (HSEs) also generalize Bayesian inference in Hilbert spaces. We investigate the link between QGMs and HSEs and show that the sum rule and Bayes rule for QGMs are equivalent to the kernel sum rule in HSEs and a special case of Nadaraya-Watson kernel regression, respectively. We show that these operations can be kernelized, and use these insights to propose a Hilbert Space Embedding of Hidden Quantum Markov Models (HSE-HQMM) to model dynamics. We present experimental results showing that HSE-HQMMs are competitive with state-of-the-art models like LSTMs and PSRNNs on several datasets, while also providing a nonparametric method for maintaining a probability distribution over continuous-valued features.

Bayesian Structure Learning by Recursive Bootstrap

Raanan Y. Rohekar, Yaniv Gurwicz, Shami Nisimov, Guy Koren, Gal Novik

We address the problem of Bayesian structure learning for domains with hundreds of variables by employing non-parametric bootstrap, recursively. We propose a method that covers both model averaging and model selection in the same framework.

The proposed method deals with the main weakness of constraint-based learning---sensitivity to errors in the independence tests---by a novel way of combining bootstrap with constraint-based learning. Essentially, we provide an algorithm for learning a tree, in which each node represents a scored CPDAG for a subset of variables and the level of the node corresponds to the maximal order of conditional independencies that are encoded in the graph. As higher order independencies are tested in deeper recursive calls, they benefit from more bootstrap samples, and therefore are more resistant to the curse-of-dimensionality. Moreover, the re-use of stable low order independencies allows greater computational efficiency. We also provide an algorithm for sampling CPDAGs efficiently from their posterior given the learned tree. That is, not from the full posterior, but from a reduced space of CPDAGs encoded in the learned tree. We empirically demonstrate that the proposed algorithm scales well to hundreds of variables, and learns better MAP models and more reliable causal relationships between variables, than other state-of-the-art-methods.

Efficient Convex Completion of Coupled Tensors using Coupled Nuclear Norms

Kishan Wimalawarne, Hiroshi Mamitsuka

Coupled norms have emerged as a convex method to solve coupled tensor completion. A limitation with coupled norms is that they only induce low-rankness using the multilinear rank of coupled tensors. In this paper, we introduce a new set of coupled norms known as coupled nuclear norms by constraining the CP rank of coupled tensors. We propose new coupled completion models using the coupled nuclear norms as regularizers, which can be optimized using computationally efficient op

timization methods. We derive excess risk bounds for proposed coupled completion models and show that proposed norms lead to better performance. Through simulation and real-data experiments, we demonstrate that proposed norms achieve better performance for coupled completion compared to existing coupled norms.

Stein Variational Gradient Descent as Moment Matching

Qiang Liu, Dilin Wang

Stein variational gradient descent (SVGD) is a non-parametric inference algorithm that evolves a set of particles to fit a given distribution of interest. We analyze the non-asymptotic properties of SVGD, showing that there exists a set of functions, which we call the Stein matching set, whose expectations are exactly estimated by any set of particles that satisfies the fixed point equation of SVGD. This set is the image of Stein operator applied on the feature maps of the positive definite kernel used in SVGD. Our results provide a theoretical framework for analyzing the properties of SVGD with different kernels, shedding insight into optimal kernel choice. In particular, we show that SVGD with linear kernels yields exact estimation of means and variances on Gaussian distributions, while random Fourier features enable probabilistic bounds for distributional approximation. Our results offer a refreshing view of the classical inference problem as fitting Stein's identity or solving the Stein equation, which may motivate more efficient algorithms.

Data-Driven Clustering via Parameterized Lloyd's Families

Maria-Florina F. 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.

Semi-Supervised Learning with Declaratively Specified Entropy Constraints

Haitian Sun, William W. Cohen, Lidong Bing

We propose a technique for declaratively specifying strategies for semi-supervised learning (SSL). SSL methods based on different assumptions perform differently on different tasks, which leads to difficulties applying them in practice. In this paper, we propose to use entropy to unify many types of constraints. Our method can be used to easily specify ensembles of semi-supervised learners, as well as agreement constraints and entropic regularization constraints between these learners, and can be used to model both well-known heuristics such as co-training, and novel domain-specific heuristics. Besides, our model is flexible as to the underlying learning mechanism. Compared to prior frameworks for specifying SSL techniques, our technique achieves consistent improvements on a suite of well-studied SSL benchmarks, and obtains a new state-of-the-art result on a difficult relation extraction task.

Bayesian Inference of Temporal Task Specifications from Demonstrations

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 demon-

nstrate 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.

Stochastic Nested Variance Reduction for Nonconvex Optimization

Dongruo Zhou, Pan Xu, Quanquan Gu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

On Markov Chain Gradient Descent

Tao Sun, Yuejiao Sun, Wotao Yin

Stochastic gradient methods are the workhorse (algorithms) of large-scale optimization problems in machine learning, signal processing, and other computational sciences and engineering. This paper studies Markov chain gradient descent, a variant of stochastic gradient descent where the random samples are taken on the trajectory of a Markov chain. Existing results of this method assume convex objectives and a reversible Markov chain and thus have their limitations. We establish new non-ergodic convergence under wider step sizes, for nonconvex problems, and for non-reversible finite-state Markov chains. Nonconvexity makes our method applicable to broader problem classes. Non-reversible finite-state Markov chains, on the other hand, can mix substantially faster. To obtain these results, we introduce a new technique that varies the mixing levels of the Markov chains. The reported numerical results validate our contributions.

The Limit Points of (Optimistic) Gradient Descent in Min-Max Optimization

Constantinos Daskalakis, Ioannis Panageas

Motivated by applications in Optimization, Game Theory, and the training of Generative Adversarial Networks, the convergence properties of first order methods in min-max problems have received extensive study. It has been recognized that they may cycle, and there is no good understanding of their limit points when they do not. When they converge, do they converge to local min-max solutions? We characterize the limit points of two basic first order methods, namely Gradient Descent/Ascent (GDA) and Optimistic Gradient Descent Ascent (OGDA). We show that both dynamics avoid unstable critical points for almost all initializations. Moreover, for small step sizes and under mild assumptions, the set of OGDA-stable critical points is a superset of GDA-stable critical points, which is a superset of local min-max solutions (strict in some cases). The connecting thread is that the behavior of these dynamics can be studied from a dynamical systems perspective.

Empirical Risk Minimization in Non-interactive Local Differential Privacy Revisited

Di Wang, Marco Gaboardi, Jinhui Xu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

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

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 cod

ing 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.

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

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 β , given by the sum of the reciprocals of the hidden layer widths. When β 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.

Learning a High Fidelity Pose Invariant Model for High-resolution Face Frontalization

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.

Provably Correct Automatic Sub-Differentiation for Qualified Programs

Sham M. Kakade, Jason D. Lee

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

CatBoost: unbiased boosting with categorical features

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin

This paper presents the key algorithmic techniques behind CatBoost, a new gradient boosting toolkit. Their combination leads to CatBoost outperforming other publicly available boosting implementations in terms of quality on a variety of datasets. Two critical algorithmic advances introduced in CatBoost are the implementation of ordered boosting, a permutation-driven alternative to the classic algorithm, and an innovative algorithm for processing categorical features. Both techniques were created to fight a prediction shift caused by a special kind of target leakage present in all currently existing implementations of gradient boosting algorithms. In this paper, we provide a detailed analysis of this problem and demonstrate that proposed algorithms solve it effectively, leading to excellent empirical results.

Constrained Generation of Semantically Valid Graphs via Regularizing Variational Autoencoders

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.

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

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 1 m and 10 m long, $105\text{ }\mu\text{m}$ diameter multimode fibre. This data is made available as the *Optical fibre inverse problem* Benchmark collection. The experimental data is used to train complex-weighted models with a range of regularisation approaches. A *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.

Scalable Hyperparameter Transfer Learning

Valerio Perrone, Rodolphe Jenatton, Matthias W. Seeger, Cedric Archambeau

Bayesian optimization (BO) is a model-based approach for gradient-free black-box function optimization, such as hyperparameter optimization. Typically, BO relies on conventional Gaussian process (GP) regression, whose algorithmic complexity is cubic in the number of evaluations. As a result, GP-based BO cannot leverage large numbers of past function evaluations, for example, to warm-start related BO runs. We propose a multi-task adaptive Bayesian linear regression model for transfer learning in BO, whose complexity is linear in the function evaluations: one Bayesian linear regression model is associated to each black-box function optimization problem (or task), while transfer learning is achieved by coupling the models through a shared deep neural net. Experiments show that the neural net learns a representation suitable for warm-starting the black-box optimization problems and that BO runs can be accelerated when the target black-box function (e.g., validation loss) is learned together with other related signals (e.g., training loss). The proposed method was found to be at least one order of magnitude faster than methods recently published in the literature.

Wasserstein Distributionally Robust Kalman Filtering

Soroosh Shafieezadeh Abadeh, Viet Anh Nguyen, Daniel Kuhn, Peyman Mohajerin Mohajerin Esfahani

We study a distributionally robust mean square error estimation problem over a n

onconvex Wasserstein ambiguity set containing only normal distributions. We show that the optimal estimator and the least favorable distribution form a Nash equilibrium. Despite the non-convex nature of the ambiguity set, we prove that the estimation problem is equivalent to a tractable convex program. We further devise a Frank-Wolfe algorithm for this convex program whose direction-searching subproblem can be solved in a quasi-closed form. Using these ingredients, we introduce a distributionally robust Kalman filter that hedges against model risk.

SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator

Cong Fang, Chris Junchi Li, Zhouchen Lin, Tong Zhang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Generalized Zero-Shot Learning with Deep Calibration Network

Shichen Liu, Mingsheng Long, Jianmin Wang, Michael I. 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.

Dual Policy Iteration

Wen Sun, Geoffrey J. Gordon, Byron Boots, J. Bagnell

Recently, a novel class of Approximate Policy Iteration (API) algorithms have demonstrated impressive practical performance (e.g., ExIt from [1], AlphaGo-Zero from [2]). This new family of algorithms maintains, and alternately optimizes, two policies: a fast, reactive policy (e.g., a deep neural network) deployed at test time, and a slow, non-reactive policy (e.g., Tree Search), that can plan multiple steps ahead. The reactive policy is updated under supervision from the non-reactive policy, while the non-reactive policy is improved with guidance from the reactive policy. In this work we study this Dual Policy Iteration (DPI) strategy in an alternating optimization framework and provide a convergence analysis that extends existing API theory. We also develop a special instance of this framework which reduces the update of non-reactive policies to model-based optimal control using learned local models, and provides a theoretically sound way of unifying model-free and model-based RL approaches with unknown dynamics. We demonstrate the efficacy of our approach on various continuous control Markov Decision Processes.

Recurrently Controlled Recurrent Networks

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 n

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.

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

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.

Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters

Pavel Dvurechenskii, Darina Dvinskikh, Alexander Gasnikov, Cesar Uribe, Angelia Nedich

We study the decentralized distributed computation of discrete approximations for the regularized Wasserstein barycenter of a finite set of continuous probability measures distributedly stored over a network. We assume there is a network of agents/machines/computers, and each agent holds a private continuous probability measure and seeks to compute the barycenter of all the measures in the network by getting samples from its local measure and exchanging information with its neighbors. Motivated by this problem, we develop, and analyze, a novel accelerated primal-dual stochastic gradient method for general stochastic convex optimization problems with linear equality constraints. Then, we apply this method to the decentralized distributed optimization setting to obtain a new algorithm for the distributed semi-discrete regularized Wasserstein barycenter problem. Moreover, we show explicit non-asymptotic complexity for the proposed algorithm. Finally, we show the effectiveness of our method on the distributed computation of the regularized Wasserstein barycenter of univariate Gaussian and von Mises distributions, as well as some applications to image aggregation.

Heterogeneous Multi-output Gaussian Process Prediction

Pablo Moreno-Muñoz, Antonio Artés, Mauricio Álvarez

We present a novel extension of multi-output Gaussian processes for handling heterogeneous outputs. We assume that each output has its own likelihood function and use a vector-valued Gaussian process prior to jointly model the parameters in all likelihoods as latent functions. Our multi-output Gaussian process uses a covariance function with a linear model of coregionalisation form. Assuming conditional independence across the underlying latent functions together with an inducing variable framework, we are able to obtain tractable variational bounds amenable to stochastic variational inference. We illustrate the performance of the model on synthetic data and two real datasets: a human behavioral study and a demographic high-dimensional dataset.

SNIPER: Efficient Multi-Scale Training

Bharat Singh, Mahyar Najibi, Larry S. 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/>.

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

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.

Delta-encoder: an effective sample synthesis method for few-shot object recognition

Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, Alex Bronstein

Learning to classify new categories based on just one or a few examples is a long-standing challenge in modern computer vision. In this work, we propose a simple yet effective method for few-shot (and one-shot) object recognition. Our approach is based on a modified auto-encoder, denoted delta-encoder, that learns to synthesize new samples for an unseen category just by seeing few examples from it. The synthesized samples are then used to train a classifier. The proposed approach learns to both extract transferable intra-class deformations, or "deltas", between same-class pairs of training examples, and to apply those deltas to the

few provided examples of a novel class (unseen during training) in order to efficiently synthesize samples from that new class. The proposed method improves the state-of-the-art of one-shot object-recognition and performs comparably in the few-shot case.

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

Peng Jiang, Gagan Agrawal

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Almost Optimal Algorithms for Linear Stochastic Bandits with Heavy-Tailed Payoffs

Han Shao, Xiaotian Yu, Irwin King, Michael R. Lyu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning towards Minimum Hyperspherical Energy

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.

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

Anqi Wu, Stan Pashkovski, Sandeep R. 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 ea

ch 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.

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

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.peihualili.org/GM-SOP>.

Neural Code Comprehension: A Learnable Representation of Code Semantics

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.

PAC-Bayes Tree: weighted subtrees with guarantees

Tin D. Nguyen, Samory Kpotufe

We present a weighted-majority classification approach over subtrees of a fixed tree, which provably achieves excess-risk of the same order as the best tree-pruning. Furthermore, the computational efficiency of pruning is maintained at both training and testing time despite having to aggregate over an exponential number

r of subtrees. We believe this is the first subtree aggregation approach with such guarantees.

Amortized Inference Regularization

Rui Shu, Hung H. 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.

Structure-Aware Convolutional Neural Networks

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.

Alternating optimization of decision trees, with application to learning sparse oblique trees

Miguel A. Carreira-Perpinan, Pooya Tavallali

Learning a decision tree from data is a difficult optimization problem. The most widespread algorithm in practice, dating to the 1980s, is based on a greedy growth of the tree structure by recursively splitting nodes, and possibly pruning back the final tree. The parameters (decision function) of an internal node are approximately estimated by minimizing an impurity measure. We give an algorithm that, given an input tree (its structure and the parameter values at its nodes), produces a new tree with the same or smaller structure but new parameter values that provably lower or leave unchanged the misclassification error. This can be applied to both axis-aligned and oblique trees and our experiments show it consistently outperforms various other algorithms while being highly scalable to large datasets and trees. Further, the same algorithm can handle a sparsity penalty, so it can learn sparse oblique trees, having a structure that is a subset of the original tree and few nonzero parameters. This combines the best of axis-aligned and oblique trees: flexibility to model correlated data, low generalization error, fast inference and interpretable nodes that involve only a few features in their decision.

Gradient Descent for Spiking Neural Networks

Dongsung Huh, Terrence J. Sejnowski

Most large-scale network models use neurons with static nonlinearities that produce analog output, despite the fact that information processing in the brain is predominantly carried out by dynamic neurons that produce discrete pulses called spikes. Research in spike-based computation has been impeded by the lack of efficient supervised learning algorithm for spiking neural networks. Here, we present a gradient descent method for optimizing spiking network models by introducing a differentiable formulation of spiking dynamics and deriving the exact gradient calculation. For demonstration, we trained recurrent spiking networks on two dynamic tasks: one that requires optimizing fast (\sim millisecond) spike-based interactions for efficient encoding of information, and a delayed-memory task over extended duration (\sim second). The results show that the gradient descent approach indeed optimizes networks dynamics on the time scale of individual spikes as well as on behavioral time scales. In conclusion, our method yields a general purpose supervised learning algorithm for spiking neural networks, which can facilitate further investigations on spike-based computations.

Statistical and Computational Trade-Offs in Kernel K-Means

Daniele Calandriello, Lorenzo Rosasco

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network

Jeffrey Pennington, Pratik Worah

An important factor contributing to the success of deep learning has been the remarkable ability to optimize large neural networks using simple first-order optimization algorithms like stochastic gradient descent. While the efficiency of such methods depends crucially on the local curvature of the loss surface, very little is actually known about how this geometry depends on network architecture and hyperparameters. In this work, we extend a recently-developed framework for studying spectra of nonlinear random matrices to characterize an important measure of curvature, namely the eigenvalues of the Fisher information matrix. We focus on a single-hidden-layer neural network with Gaussian data and weights and provide an exact expression for the spectrum in the limit of infinite width. We find that linear networks suffer worse conditioning than nonlinear networks and that nonlinear networks are generically non-degenerate. We also predict and demonstrate empirically that by adjusting the nonlinearity, the spectrum can be tuned so as to improve the efficiency of first-order optimization methods.

Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks

Grant Rotskoff, Eric Vanden-Eijnden

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Uplift Modeling from Separate Labels

Ikko Yamane, Florian Yger, Jamal Atif, Masashi Sugiyama

Uplift modeling is aimed at estimating the incremental impact of an action on an individual's behavior, which is useful in various application domains such as targeted marketing (advertisement campaigns) and personalized medicine (medical treatments). Conventional methods of uplift modeling require every instance to be jointly equipped with two types of labels: the taken action and its outcome. However, obtaining two labels for each instance at the same time is difficult or expensive in many real-world problems. In this paper, we propose a novel method o

f uplift modeling that is applicable to a more practical setting where only one type of labels is available for each instance. We show a mean squared error bound for the proposed estimator and demonstrate its effectiveness through experiments.

A Bridging Framework for Model Optimization and Deep Propagation

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.

Learning filter widths of spectral decompositions with wavelets

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.

Learning a Warping Distance from Unlabeled Time Series Using Sequence Autoencoders

Abubakar Abid, James Y. 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.

ories. 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.

Gen-Oja: Simple & Efficient Algorithm for Streaming Generalized Eigenvector Computation

Kush Bhatia, Aldo Pacchiano, Nicolas Flammarion, Peter L. Bartlett, Michael I. Jordan

In this paper, we study the problems of principle Generalized Eigenvector computation and Canonical Correlation Analysis in the stochastic setting. We propose a simple and efficient algorithm for these problems. We prove the global convergence of our algorithm, borrowing ideas from the theory of fast-mixing Markov chains and two-Time-Scale Stochastic Approximation, showing that it achieves the optimal rate of convergence. In the process, we develop tools for understanding stochastic processes with Markovian noise which might be of independent interest.

Heterogeneous Bitwidth Binarization in Convolutional Neural Networks

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.

BRUNO: A Deep Recurrent Model for Exchangeable Data

Iryna Korshunova, Jonas Degraeve, Ferenc Huszar, Yarin Gal, Arthur Gretton, Joni Dambre

We present a novel model architecture which leverages deep learning tools to perform exact Bayesian inference on sets of high dimensional, complex observations.

Our model is provably exchangeable, meaning that the joint distribution over observations is invariant under permutation: this property lies at the heart of Bayesian inference. The model does not require variational approximations to train, and new samples can be generated conditional on previous samples, with cost linear in the size of the conditioning set. The advantages of our architecture are demonstrated on learning tasks that require generalisation from short observed sequences while modelling sequence variability, such as conditional image generation, few-shot learning, and anomaly detection.

Asymptotic optimality of adaptive importance sampling

François Portier, Bernard Delyon

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Phase Retrieval Under a Generative Prior

Paul Hand, Oscar Leong, Vlad Voroninski

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Gaussian Process Prior Variational Autoencoders

Francesco Paolo Casale, Adrian Dalca, Luca Saglietti, Jennifer Listgarten, Nicolo Fusi

Variational autoencoders (VAE) are a powerful and widely-used class of models to learn complex data distributions in an unsupervised fashion. One important limitation of VAEs is the prior assumption that latent sample representations are independent and identically distributed. However, for many important datasets, such as time-series of images, this assumption is too strong: accounting for covariances between samples, such as those in time, can yield to a more appropriate model specification and improve performance in downstream tasks. In this work, we introduce a new model, the Gaussian Process (GP) Prior Variational Autoencoder (GPPVAE), to specifically address this issue. The GPPVAE aims to combine the power of VAEs with the ability to model correlations afforded by GP priors. To achieve efficient inference in this new class of models, we leverage structure in the covariance matrix, and introduce a new stochastic backpropagation strategy that allows for computing stochastic gradients in a distributed and low-memory fashion. We show that our method outperforms conditional VAEs (CVAEs) and an adaptation of standard VAEs in two image data applications.

Deep Predictive Coding Network with Local Recurrent Processing for Object Recognition

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 non-linear 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.

Exponentiated Strongly Rayleigh Distributions

Zelda E. Mariet, Suvrit Sra, Stefanie Jegelka

Strongly Rayleigh (SR) measures are discrete probability distributions over the subsets of a ground set. They enjoy strong negative dependence properties, as a result of which they assign higher probability to subsets of diverse elements. We introduce in this paper Exponentiated Strongly Rayleigh (ESR) measures, which sharpen (or smoothen) the negative dependence property of SR measures via a single parameter (the exponent) that can intuitively be understood as an inverse temperature. We develop efficient MCMC procedures for approximate sampling from ESRs, and obtain explicit mixing time bounds for two concrete instances: exponentiated versions of Determinantal Point Processes and Dual Volume Sampling. We illustrate some of the potential of ESRs, by applying them to a few machine learning tasks; empirical results confirm that beyond their theoretical appeal, ESR-based models hold significant promise for these tasks.

Variational Inference with Tail-adaptive f-Divergence

Dilin Wang, Hao Liu, Qiang Liu

Variational inference with α -divergences has been widely used in modern probabilistic

machine learning. Compared to Kullback-Leibler (KL) divergence, a major advantage of using α -divergences (with positive α values) is their mass-covering property. However, estimating and optimizing α -divergences require to use importance

sampling, which could have extremely large or infinite variances due to heavy tails of importance weights. In this paper, we propose a new class of tail-adaptive f-divergences that adaptively change the convex function f with the

tail of the importance weights, in a way that theoretically guarantee finite moments,

while simultaneously achieving mass-covering properties. We test our methods on Bayesian neural networks, as well as deep reinforcement learning in which our method is applied to improve a recent soft actor-critic (SAC) algorithm (Haarnoja

et al., 2018). Our results show that our approach yields significant advantages compared with existing methods based on classical KL and α -divergences.

Modelling and unsupervised learning of symmetric deformable object categories

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.

Generalizing to Unseen Domains via Adversarial Data Augmentation

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, Silvio Savarese

We are concerned with learning models that generalize well to different unseen domains. We consider a worst-case formulation over data distributions that are near the source domain in the feature space. Only using training data from a single

source distribution, we propose an iterative procedure that augments the dataset with examples from a fictitious target domain that is "hard" under the current model. We show that our iterative scheme is an adaptive data augmentation method where we append adversarial examples at each iteration. For softmax losses, we show that our method is a data-dependent regularization scheme that behaves differently from classical regularizers that regularize towards zero (e.g., ridge or lasso). On digit recognition and semantic segmentation tasks, our method learns models improve performance across a range of a priori unknown target domains.

Information-theoretic Limits for Community Detection in Network Models

Chuyang Ke, Jean Honorio

We analyze the information-theoretic limits for the recovery of node labels in several network models. This includes the Stochastic Block Model, the Exponential Random Graph Model, the Latent Space Model, the Directed Preferential Attachment

t Model, and the Directed Small-world Model. For the Stochastic Block Model, the non-recoverability condition depends on the probabilities of having edges inside a community, and between different communities. For the Latent Space Model, the non-recoverability condition depends on the dimension of the latent space, and how far and spread are the communities in the latent space. For the Directed Preferential Attachment Model and the Directed Small-world Model, the non-recoverability condition depends on the ratio between homophily and neighborhood size. We also consider dynamic versions of the Stochastic Block Model and the Latent Space Model.

Dendritic cortical microcircuits approximate the backpropagation algorithm

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.

Structured Local Minima in Sparse Blind Deconvolution

Yugian Zhang, Han-wen Kuo, John Wright

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Solving Non-smooth Constrained Programs with Lower Complexity than $\mathcal{O}(1/\epsilon)$: A Primal-Dual Homotopy Smoothing Approach

Xiaohan Wei, Hao Yu, Qing Ling, Michael Neely

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces

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-t

o-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.

Isolating Sources of Disentanglement in Variational Autoencoders

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, David K. Duvenaud

We decompose the evidence lower bound to show the existence of a term measuring the total correlation between latent variables. We use this to motivate the beta-TCVAE (Total Correlation Variational Autoencoder) algorithm, a refinement and plug-in replacement of the beta-VAE for learning disentangled representations, requiring no additional hyperparameters during training. We further propose a principled classifier-free measure of disentanglement called the mutual information gap (MIG). We perform extensive quantitative and qualitative experiments, in both restricted and non-restricted settings, and show a strong relation between total correlation and disentanglement, when the model is trained using our framework.

Learning to Share and Hide Intentions using Information Regularization

DJ Strouse, Max Kleiman-Weiner, Josh Tenenbaum, Matt Botvinick, David J. Schwab

Learning to cooperate with friends and compete with foes is a key component of multi-agent reinforcement learning. Typically to do so, one requires access to either a model of or interaction with the other agent(s). Here we show how to learn effective strategies for cooperation and competition in an asymmetric information game with no such model or interaction. Our approach is to encourage an agent to reveal or hide their intentions using an information-theoretic regularizer. We consider both the mutual information between goal and action given state, as well as the mutual information between goal and state. We show how to stochastically optimize these regularizers in a way that is easy to integrate with policy gradient reinforcement learning. Finally, we demonstrate that cooperative (competitive) policies learned with our approach lead to more (less) reward for a second agent in two simple asymmetric information games.

Why Is My Classifier Discriminatory?

Irene Chen, Fredrik D. Johansson, David Sontag

Recent attempts to achieve fairness in predictive models focus on the balance between fairness and accuracy. In sensitive applications such as healthcare or criminal justice, this trade-off is often undesirable as any increase in prediction error could have devastating consequences. In this work, we argue that the fairness of predictions should be evaluated in context of the data, and that unfairness induced by inadequate sample sizes or unmeasured predictive variables should be addressed through data collection, rather than by constraining the model. We decompose cost-based metrics of discrimination into bias, variance, and noise, and propose actions aimed at estimating and reducing each term. Finally, we perform case-studies on prediction of income, mortality, and review ratings, confirming the value of this analysis. We find that data collection is often a means to reduce discrimination without sacrificing accuracy.

Unsupervised Learning of Object Landmarks through Conditional Image Generation

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 distills 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 poi

nt 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.

Scalable Coordinated Exploration in Concurrent Reinforcement Learning

Maria Dimakopoulou, Ian Osband, Benjamin Van Roy

We consider a team of reinforcement learning agents that concurrently operate in a common environment, and we develop an approach to efficient coordinated exploration that is suitable for problems of practical scale. Our approach builds on the seed sampling concept introduced in Dimakopoulou and Van Roy (2018) and on a randomized value function learning algorithm from Osband et al. (2016). We demonstrate that, for simple tabular contexts, the approach is competitive with those previously proposed in Dimakopoulou and Van Roy (2018) and with a higher-dimensional problem and a neural network value function representation, the approach learns quickly with far fewer agents than alternative exploration schemes.

Bayesian Semi-supervised Learning with Graph Gaussian Processes

Yin Cheng Ng, Nicolò Colombo, Ricardo Silva

We propose a data-efficient Gaussian process-based Bayesian approach to the semi-supervised learning problem on graphs. The proposed model shows extremely competitive performance when compared to the state-of-the-art graph neural networks on semi-supervised learning benchmark experiments, and outperforms the neural networks in active learning experiments where labels are scarce. Furthermore, the model does not require a validation data set for early stopping to control overfitting. Our model can be viewed as an instance of empirical distribution regression weighted locally by network connectivity. We further motivate the intuitive construction of the model with a Bayesian linear model interpretation where the node features are filtered by an operator related to the graph Laplacian. The method can be easily implemented by adapting off-the-shelf scalable variational inference algorithms for Gaussian processes.

Simple, Distributed, and Accelerated Probabilistic Programming

Dustin Tran, Matthew W. Hoffman, Dave Moore, Christopher Suter, Srinivas Vasudevan, Alexey Radul

We describe a simple, low-level approach for embedding probabilistic programming in a deep learning ecosystem. In particular, we distill probabilistic programming down to a single abstraction—the random variable. Our lightweight implementation in TensorFlow enables numerous applications: a model-parallel variational auto-encoder (VAE) with 2nd-generation tensor processing units (TPUV2s); a data-parallel autoregressive model (Image Transformer) with TPUV2s; and multi-GPU No-U-Turn Sampler (NUTS). For both a state-of-the-art VAE on 64x64 ImageNet and Image Transformer on 256x256 CelebA-HQ, our approach achieves an optimal linear speedup from 1 to 256 TPUV2 chips. With NUTS, we see a 100x speedup on GPUs over Stan and 37x over PyMC3.

When do random forests fail?

Cheng Tang, Damien Garreau, Ulrike von Luxburg

Random forests are learning algorithms that build large collections of random trees and make predictions by averaging the individual tree predictions.

In this paper, we consider various tree constructions and examine how the choice of parameters affects the generalization error of the resulting random forests as the sample size goes to infinity.

We show that subsampling of data points during the tree construction phase is important: Forests can become inconsistent with either no subsampling or too severe subsampling.

As a consequence, even highly randomized trees can lead to inconsistent forests if no subsampling is used, which implies that some of the commonly used setups f

or random forests can be inconsistent.

As a second consequence we can show that trees that have good performance in nearest-neighbor search can be a poor choice for random forests.

Contextual Combinatorial Multi-armed Bandits with Volatile Arms and Submodular Reward

Lixing Chen, Jie Xu, Zhuo Lu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning to Reconstruct Shapes from Unseen Classes

Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, Jiajun Wu

From a single image, humans are able to perceive the full 3D shape of an object by exploiting learned shape priors from everyday life. Contemporary single-image 3D reconstruction algorithms aim to solve this task in a similar fashion, but often end up with priors that are highly biased by training classes. Here we present an algorithm, Generalizable Reconstruction (GenRe), designed to capture more generic, class-agnostic shape priors. We achieve this with an inference network and training procedure that combine 2.5D representations of visible surfaces (depth and silhouette), spherical shape representations of both visible and non-visible surfaces, and 3D voxel-based representations, in a principled manner that exploits the causal structure of how 3D shapes give rise to 2D images. Experiments demonstrate that GenRe performs well on single-view shape reconstruction, and generalizes to diverse novel objects from categories not seen during training.

Mixture Matrix Completion

Daniel Pimentel-Alarcon

Completing a data matrix X has become an ubiquitous problem in modern data science, with motivations in recommender systems, computer vision, and networks inference, to name a few. One typical assumption is that X is low-rank. A more general model assumes that each column of X corresponds to one of several low-rank matrices. This paper generalizes these models to what we call mixture matrix completion (MMC): the case where each entry of X corresponds to one of several low-rank matrices. MMC is a more accurate model for recommender systems, and brings more flexibility to other completion and clustering problems. We make four fundamental contributions about this new model. First, we show that MMC is theoretically possible (well-posed). Second, we give its precise information-theoretic identifiability conditions. Third, we derive the sample complexity of MMC. Finally, we give a practical algorithm for MMC with performance comparable to the state-of-the-art for simpler related problems, both on synthetic and real data.

Fighting Boredom in Recommender Systems with Linear Reinforcement Learning

Romain WARLOP, Alessandro Lazaric, Jérémie Mary

A common assumption in recommender systems (RS) is the existence of a best fixed recommendation strategy. Such strategy may be simple and work at the item level (e.g., in multi-armed bandit it is assumed one best fixed arm/item exists) or implement more sophisticated RS (e.g., the objective of A/B testing is to find the

best fixed RS and execute it thereafter). We argue that this assumption is rarely verified in practice, as the recommendation process itself may impact the user's

preferences. For instance, a user may get bored by a strategy, while she may gain interest again, if enough time passed since the last time that strategy was used. In

this case, a better approach consists in alternating different solutions at the right frequency to fully exploit their potential. In this paper, we first cast the problem as

a Markov decision process, where the rewards are a linear function of the recent history of actions, and we show that a policy considering the long-term influence of the recommendations may outperform both fixed-action and contextual greedy policies. We then introduce an extension of the UCRL algorithm (L IN UCRL) to effectively balance exploration and exploitation in an unknown environment, and we derive a regret bound that is independent of the number of states. Finally, we empirically validate the model assumptions and the algorithm in a number of realistic scenarios.

Diffusion Maps for Textual Network Embedding

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.

Out-of-Distribution Detection using Multiple Semantic Label Representations

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.

First-order Stochastic Algorithms for Escaping From Saddle Points in Almost Linear Time

Yi Xu, Rong Jin, Tianbao Yang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

cpSGD: Communication-efficient and differentially-private distributed SGD

Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X. Yu, Sanjiv Kumar, Brendan McMahan

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Factored Bandits

Julian Zimmert, Yevgeny Seldin

We introduce the factored bandits model, which is a framework for learning with limited (bandit) feedback, where actions can be decomposed into a Cartesian

product of atomic actions. Factored bandits incorporate rank-1 bandits as a special case, but significantly relax the assumptions on the form of the reward function. We provide an anytime algorithm for stochastic factored bandits and up to constants matching upper and lower regret bounds for the problem. Furthermore, we show that with a slight modification the proposed algorithm can be applied to utility based dueling bandits. We obtain an improvement in the additive terms of the regret bound compared to state of the art algorithms (the additive terms are dominating up to time horizons which are exponential in the number of arms).

Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suci, Christoph Studer, Tudor Dumitras, Tom Goldstein

Data poisoning is an attack on machine learning models wherein the attacker adds examples to the training set to manipulate the behavior of the model at test time. This paper explores poisoning attacks on neural nets. The proposed attacks use ``clean-labels''; they don't require the attacker to have any control over the labeling of training data. They are also targeted; they control the behavior of the classifier on a specific test instance without degrading overall classifier performance. For example, an attacker could add a seemingly innocuous image (that is properly labeled) to a training set for a face recognition engine, and control the identity of a chosen person at test time. Because the attacker does not need to control the labeling function, poisons could be entered into the training set simply by putting them online and waiting for them to be scraped by a data collection bot.

Implicit Probabilistic Integrators for ODEs

Onur Teymur, Han Cheng Lie, Tim Sullivan, Ben Calderhead

We introduce a family of implicit probabilistic integrators for initial value problems (IVPs), taking as a starting point the multistep Adams-Moulton method. The implicit construction allows for dynamic feedback from the forthcoming time-step, in contrast to previous probabilistic integrators, all of which are based on explicit methods. We begin with a concise survey of the rapidly-expanding field of probabilistic ODE solvers. We then introduce our method, which builds on and adapts the work of Conrad et al. (2016) and Teymur et al. (2016), and provide a rigorous proof of its well-definedness and convergence. We discuss the problem of the calibration of such integrators and suggest one approach. We give an illustrative example highlighting the effect of the use of probabilistic integrators—including our new method—in the setting of parameter inference within an inverse problem.

Non-metric Similarity Graphs for Maximum Inner Product Search

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.

Learning convex polytopes with margin

Lee-Ad Gottlieb, Eran Kaufman, Aryeh Kontorovich, Gabriel Nivasch

We present improved algorithm for properly learning convex polytopes in the realizable PAC setting from data with a margin. Our learning algorithm constructs

a consistent polytope as an intersection of about $t \log t$ halfspaces with margin s

in time polynomial in t (where t is the number of halfspaces forming an optimal polytope).

We also identify distinct generalizations of the notion of margin from hyperplanes

to polytopes and investigate how they relate geometrically; this result may be of

interest beyond the learning setting.

Distilled Wasserstein Learning for Word Embedding and Topic Modeling

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.

Inferring Latent Velocities from Weather Radar Data using Gaussian Processes

Rico Angell, Daniel R. 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.

Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization

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 dist

tributional 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.

Multi-Agent Generative Adversarial Imitation Learning

Jiaming Song, Hongyu Ren, Dorsa Sadigh, Stefano Ermon

Imitation learning algorithms can be used to learn a policy from expert demonstrations without access to a reward signal. However, most existing approaches are not applicable in multi-agent settings due to the existence of multiple (Nash) equilibria and non-stationary environments.

We propose a new framework for multi-agent imitation learning for general Markov games, where we build upon a generalized notion of inverse reinforcement learning. We further introduce a practical multi-agent actor-critic algorithm with good empirical performance. Our method can be used to imitate complex behaviors in high-dimensional environments with multiple cooperative or competing agents.

Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning

Supasorn Suwajanakorn, Noah Snavely, Jonathan J. Tompson, Mohammad Norouzi

This paper presents KeypointNet, an end-to-end geometric reasoning framework to learn an optimal set of category-specific keypoints, along with their detectors to predict 3D keypoints in a single 2D input image. We demonstrate this framework on 3D pose estimation task by proposing a differentiable pose objective that seeks the optimal set of keypoints for recovering the relative pose between two views of an object. Our network automatically discovers a consistent set of keypoints across viewpoints of a single object as well as across all object instances of a given object class. Importantly, we find that our end-to-end approach using no ground-truth keypoint annotations outperforms a fully supervised baseline using the same neural network architecture for the pose estimation task.

The discovered 3D keypoints across the car, chair, and plane categories of ShapeNet are visualized at <https://keypoints.github.io/>

Variational Learning on Aggregate Outputs with Gaussian Processes

Ho Chung Law, Dino Sejdinovic, Ewan Cameron, Tim Lucas, Seth Flaxman, Katherine Battle, Kenji Fukumizu

While a typical supervised learning framework assumes that the inputs and the outputs are measured at the same levels of granularity, many applications, including global mapping of disease, only have access to outputs at a much coarser level than that of the inputs. Aggregation of outputs makes generalization to new inputs much more difficult. We consider an approach to this problem based on variational learning with a model of output aggregation and Gaussian processes, where aggregation leads to intractability of the standard evidence lower bounds. We propose new bounds and tractable approximations, leading to improved prediction accuracy and scalability to large datasets, while explicitly taking uncertainty into account. We develop a framework which extends to several types of likelihoods, including the Poisson model for aggregated count data. We apply our framework to a challenging and important problem, the fine-scale spatial modelling of malaria incidence, with over 1 million observations.

Adaptation to Easy Data in Prediction with Limited Advice

Tobias Sommer Thune, Yevgeny Seldin

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Maximum Causal Tsallis Entropy Imitation Learning

Kyungjae Lee, Sungjoon Choi, Songhwai Oh

In this paper, we propose a novel maximum causal Tsallis entropy (MCTE) framework

k for imitation learning which can efficiently learn a sparse multi-modal policy distribution from demonstrations. We provide the full mathematical analysis of the proposed framework. First, the optimal solution of an MCTE problem is shown to be a sparsemax distribution, whose supporting set can be adjusted. The proposed method has advantages over a softmax distribution in that it can exclude unnecessary actions by assigning zero probability. Second, we prove that an MCTE problem is equivalent to robust Bayes estimation in the sense of the Brier score. Third, we propose a maximum causal Tsallis entropy imitation learning (MCTEIL) algorithm with a sparse mixture density network (sparse MDN) by modeling mixture weights using a sparsemax distribution. In particular, we show that the causal Tsallis entropy of an MDN encourages exploration and efficient mixture utilization while Boltzmann Gibbs entropy is less effective. We validate the proposed method in two simulation studies and MCTEIL outperforms existing imitation learning methods in terms of average returns and learning multi-modal policies.

Importance Weighting and Variational Inference

Justin Domke, Daniel R. Sheldon

Recent work used importance sampling ideas for better variational bounds on likelihoods. We clarify the applicability of these ideas to pure probabilistic inference, by showing the resulting Importance Weighted Variational Inference (IWVI) technique is an instance of augmented variational inference, thus identifying the looseness in previous work. Experiments confirm IWVI's practicality for probabilistic inference. As a second contribution, we investigate inference with elliptical distributions, which improves accuracy in low dimensions, and convergence in high dimensions.

Mapping Images to Scene Graphs with Permutation-Invariant Structured Prediction

Roee Herzig, Moshiko Raboh, Gal Chechik, Jonathan Berant, Amir Globerson

Machine understanding of complex images is a key goal of artificial intelligence. One challenge underlying this task is that visual scenes contain multiple inter-related objects, and that global context plays an important role in interpreting the scene. A natural modeling framework for capturing such effects is structured prediction, which optimizes over complex labels, while modeling within-label interactions. However, it is unclear what principles should guide the design of a structured prediction model that utilizes the power of deep learning components. Here we propose a design principle for such architectures that follows from a natural requirement of permutation invariance. We prove a necessary and sufficient characterization for architectures that follow this invariance, and discuss its implication on model design. Finally, we show that the resulting model achieves new state of the art results on the Visual Genome scene graph labeling benchmark, outperforming all recent approaches.

Are ResNets Provably Better than Linear Predictors?

Ohad Shamir

A residual network (or ResNet) is a standard deep neural net architecture, with state-of-the-art performance across numerous applications. The main premise of ResNets is that they allow the training of each layer to focus on fitting just the residual of the previous layer's output and the target output. Thus, we should expect that the trained network is no worse than what we can obtain if we remove the residual layers and train a shallower network instead. However, due to the non-convexity of the optimization problem, it is not at all clear that ResNets indeed achieve this behavior, rather than getting stuck at some arbitrarily poor local minimum. In this paper, we rigorously prove that arbitrarily deep, nonlinear residual units indeed exhibit this behavior, in the sense that the optimization landscape contains no local minima with value above what can be obtained with a linear predictor (namely a 1-layer network). Notably, we show this under minimal or no assumptions on the precise network architecture, data distribution, or loss function used. We also provide a quantitative analysis of approximate stationary points for this problem. Finally, we show that with a certain tweak to the architecture, training the network with standard stochastic gradient descent

achieves an objective value close or better than any linear predictor.

Meta-Gradient Reinforcement Learning

Zhongwen Xu, Hado P. van Hasselt, David Silver

The goal of reinforcement learning algorithms is to estimate and/or optimise the value function. However, unlike supervised learning, no teacher or oracle is available to provide the true value function. Instead, the majority of reinforcement

learning algorithms estimate and/or optimise a proxy for the value function. This

proxy is typically based on a sampled and bootstrapped approximation to the true value function, known as a return. The particular choice of return is one of the chief components determining the nature of the algorithm: the rate at which future

rewards are discounted; when and how values should be bootstrapped; or even the nature of the rewards themselves. It is well-known that these decisions are crucial

to the overall success of RL algorithms. We discuss a gradient-based meta-learning

algorithm that is able to adapt the nature of the return, online, whilst interacting

and learning from the environment. When applied to 57 games on the Atari 2600 environment over 200 million frames, our algorithm achieved a new state-of-the-art

performance.

GPYtorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration

Jacob Gardner, Geoff Pleiss, Kilian Q. Weinberger, David Bindel, Andrew G. Wilson

Despite advances in scalable models, the inference tools used for Gaussian processes (GPs) have yet to fully capitalize on developments in computing hardware. We present an efficient and general approach to GP inference based on Blackbox Matrix-Matrix multiplication (BBMM). BBMM inference uses a modified batched version of the conjugate gradients algorithm to derive all terms for training and inference in a single call. BBMM reduces the asymptotic complexity of exact GP inference from $O(n^3)$ to $O(n^2)$. Adapting this algorithm to scalable approximations and complex GP models simply requires a routine for efficient matrix-matrix multiplication with the kernel and its derivative. In addition, BBMM uses a specialized preconditioner to substantially speed up convergence. In experiments we show that BBMM effectively uses GPU hardware to dramatically accelerate both exact GP inference and scalable approximations. Additionally, we provide GPYtorch, a software platform for scalable GP inference via BBMM, built on PyTorch.

Spectral Signatures in Backdoor Attacks

Brandon Tran, Jerry Li, Aleksander Madry

A recent line of work has uncovered a new form of data poisoning: so-called backdoor attacks. These attacks are particularly dangerous because they do not affect a network's behavior on typical, benign data. Rather, the network only deviates from its expected output when triggered by an adversary's planted perturbation.

Uncertainty-Aware Attention for Reliable Interpretation and Prediction

Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, Sung Ju Hwang

Attention mechanism is effective in both focusing the deep learning models on relevant features and interpreting them. However, attentions may be unreliable since the networks that generate them are often trained in a weakly-supervised manner. To overcome this limitation, we introduce the notion of input-dependent uncertainty to the attention mechanism, such that it generates attention for each fe

ature with varying degrees of noise based on the given input, to learn larger variance on instances it is uncertain about. We learn this Uncertainty-aware Attention (UA) mechanism using variational inference, and validate it on various risk prediction tasks from electronic health records on which our model significantly outperforms existing attention models. The analysis of the learned attentions shows that our model generates attentions that comply with clinicians' interpretation, and provide richer interpretation via learned variance. Further evaluation of both the accuracy of the uncertainty calibration and the prediction performance with "I don't know" decision show that UA yields networks with high reliability as well.

Attention in Convolutional LSTM for Gesture Recognition

Liang Zhang, Guangming Zhu, Lin Mei, Peiyi Shen, Syed Afaq Ali Shah, Mohammed Benmamoun

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.

Forward Modeling for Partial Observation Strategy Games - A StarCraft Defogger

Gabriel Synnaeve, Zeming Lin, Jonas Gehring, Dan Gant, Vegard Mella, Vasil Khali

dov, Nicolas Carion, Nicolas Usunier

We formulate the problem of defogging as state estimation and future state prediction from previous, partial observations in the context of real-time strategy games. We propose to employ encoder-decoder neural networks for this task, and introduce proxy tasks and baselines for evaluation to assess their ability of capturing basic game rules and high-level dynamics. By combining convolutional neural networks and recurrent networks, we exploit spatial and sequential correlations and train well-performing models on a large dataset of human games of StarCraft: Brood War. Finally, we demonstrate the relevance of our models to downstream tasks by applying them for enemy unit prediction in a state-of-the-art, rule-based StarCraft bot. We observe improvements in win rates against several strong community bots.

PacGAN: The power of two samples in generative adversarial networks

Zinan Lin, Ashish Kheta, 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.

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

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.

Multilingual Anchoring: Interactive Topic Modeling and Alignment Across Languages

Michelle Yuan, Benjamin Van Durme, Jordan L. Ying

Multilingual topic models can reveal patterns in cross-lingual document collections. However, existing models lack speed and interactivity, which prevents adoption in everyday corpora exploration or quick moving situations (e.g., natural disasters, political instability). First, we propose a multilingual anchoring algorithm that builds an anchor-based topic model for documents in different languages. Then, we incorporate interactivity to develop MTAnchor (Multilingual Topic Anchors), a system that allows users to refine the topic model. We test our algorithms on labeled English, Chinese, and Sinhalese documents. Within minutes, our methods can produce interpretable topics that are useful for specific classification tasks.

Generalized Inverse Optimization through Online Learning

Chaosheng Dong, Yiran Chen, Bo Zeng

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Sanity Checks for Saliency Maps

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Bee N Kim

Saliency methods have emerged as a popular tool to highlight features in an input

deemed relevant for the prediction of a learned model. Several saliency methods have been proposed, often guided by visual appeal on image data. In this work, we

propose an actionable methodology to evaluate what kinds of explanations a given method can and cannot provide. We find that reliance, solely, on visual assessment

can be misleading. Through extensive experiments we show that some existing

saliency methods are independent both of the model and of the data generating process. Consequently, methods that fail the proposed tests are inadequate for tasks that are sensitive to either data or model, such as, finding outliers in the data,

explaining the relationship between inputs and outputs that the model learned, and debugging the model. We interpret our findings through an analogy with edge detection in images, a technique that requires neither training data nor model.

Theory in the case of a linear model and a single-layer convolutional neural network supports our experimental findings.

Differentially Private Uniformly Most Powerful Tests for Binomial Data

Jordan Awan, Aleksandra Slavković

We derive uniformly most powerful (UMP) tests for simple and one-sided hypotheses for a population proportion within the framework of Differential Privacy (DP), optimizing finite sample performance. We show that in general, DP hypothesis tests can be written in terms of linear constraints, and for exchangeable data can always be expressed as a function of the empirical distribution. Using this structure, we prove a 'Neyman-Pearson lemma' for binomial data under DP, where the DP-UMP only depends on the sample sum. Our tests can also be stated as a post-processing of a random variable, whose distribution we coin "Truncated-Uniform-Laplace" (Tulap), a generalization of the Staircase and discrete Laplace distributions. Furthermore, we obtain exact p-values, which are easily computed in terms of the Tulap random variable. We show that our results also apply to distribution-free hypothesis tests for continuous data. Our simulation results demonstrate that our tests have exact type I error, and are more powerful than current techniques.

Bayesian Alignments of Warped Multi-Output Gaussian Processes

Markus Kaiser, Clemens Otte, Thomas Runkler, Carl Henrik Ek

We propose a novel Bayesian approach to modelling nonlinear alignments of time series based on latent shared information. We apply the method to the real-world problem of finding common structure in the sensor data of wind turbines introduced by the underlying latent and turbulent wind field. The proposed model allows for both arbitrary alignments of the inputs and non-parametric output warpings to transform the observations. This gives rise to multiple deep Gaussian process models connected via latent generating processes. We present an efficient variational approximation based on nested variational compression and show how the model can be used to extract shared information between dependent time series, recovering an interpretable functional decomposition of the learning problem. We show results for an artificial data set and real-world data of two wind turbines.

Semidefinite relaxations for certifying robustness to adversarial examples

Aditi Raghunathan, Jacob Steinhardt, Percy S. 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.

Compact Representation of Uncertainty in Clustering

Craig Greenberg, Nicholas Monath, Ari Kobren, Patrick Flaherty, Andrew McGregor, Andrew McCallum

For many classic structured prediction problems, probability distributions over the dependent variables can be efficiently computed using widely-known algorithms and data structures (such as forward-backward, and its corresponding trellis for exact probability distributions in Markov models). However, we know of no previous work studying efficient representations of exact distributions over clusterings. This paper presents definitions and proofs for a dynamic-programming inference procedure that computes the partition function, the marginal probability of a cluster, and the MAP clustering---all exactly. Rather than the N^{th} Bell number, these exact solutions take time and space proportional to the substantially smaller powerset of N . Indeed, we improve upon the time complexity of the algorithm introduced by Kohonen and Corander (2016) for this problem by a factor of N . While still large, this previously unknown result is intellectually interesting in its own right, makes feasible exact inference for important real-world small data applications (such as medicine), and provides a natural stepping stone towards sparse-trellis approximations that enable further scalability (which we also explore). In experiments, we demonstrate the superiority of our approach over approximate methods in analyzing real-world gene expression data used in cancer treatment.

DeepPINK: reproducible feature selection in deep neural networks

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.

Supervised autoencoders: Improving generalization performance with unsupervised regularizers

Lei Le, Andrew Patterson, Martha White

Generalization performance is a central goal in machine learning, particularly when learning representations with large neural networks. A common strategy to improve generalization has been through the use of regularizers, typically as a norm constraining the parameters. Regularizing hidden layers in a neural network architecture, however, is not straightforward. There have been a few effective layer-wise suggestions, but without theoretical guarantees for improved performance. In this work, we theoretically and empirically analyze one such model, called a supervised auto-encoder: a neural network that predicts both inputs (reconstruction error) and targets jointly. We provide a novel generalization result for linear auto-encoders, proving uniform stability based on the inclusion of the reconstruction error---particularly as an improvement on simplistic regularization such as norms or even on more advanced regularizations such as the use of auxiliary tasks. Empirically, we then demonstrate that, across an array of architectures with a different number of hidden units and activation functions, the supervised auto-encoder compared to the corresponding standard neural network never harms performance and can significantly improve generalization.

Understanding Regularized Spectral Clustering via Graph Conductance

Yilin Zhang, Karl Rohe

This paper uses the relationship between graph conductance and spectral clustering to study (i) the failures of spectral clustering and (ii) the benefits of regularization. The explanation is simple. Sparse and stochastic graphs create several dangling sets'', or small trees that are connected to the core of the graph by only one edge. Graph conductance is sensitive to these noisy dangling sets and spectral clustering inherits this sensitivity. The second part of the paper starts from a previously proposed form of regularized spectral clustering and shows that it is related to the graph conductance on a regularized graph''. When graph conductance is computed on the regularized graph, we call it CoreCut. Based upon previous arguments that relate graph conductance to spectral clustering (e.g. Cheeger inequality), minimizing CoreCut relaxes to regularized spectral clustering. Simple inspection of CoreCut reveals why it is less sensitive to dangling sets. Together, these results show that unbalanced partitions from spectral clustering can be understood as overfitting to noise in the periphery of a sparse and stochastic graph. Regularization fixes this overfitting. In addition to this statistical benefit, these results also demonstrate how regularization can improve the computational speed of spectral clustering. We provide simulations and data examples to illustrate these results.

Plug-in Estimation in High-Dimensional Linear Inverse Problems: A Rigorous Analysis

Alyson K. Fletcher, Parthe Pandit, Sundeep Rangan, Subrata Sarkar, Philip Schnitger

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Multitask Boosting for Survival Analysis with Competing Risks

Alexis Bellot, Mihaela van der Schaar

The co-occurrence of multiple diseases among the general population is an important problem as those patients have more risk of complications and represent a large share of health care expenditure. Learning to predict time-to-event probabilities for these patients is a challenging problem because the risks of events are correlated (there are competing risks) with often only few patients experiencing individual events of interest, and of those only a fraction are actually observed in the data. We introduce in this paper a survival model with the flexibility to leverage a common representation of related events that is designed to correct for the strong imbalance in observed outcomes. The procedure is sequential: outcome-specific survival distributions form the components of nonparametric multivariate estimators which we combine into an ensemble in such a way as to ensure accurate predictions on all outcome types simultaneously. Our algorithm is general and represents the first boosting-like method for time-to-event data with multiple outcomes. We demonstrate the performance of our algorithm on synthetic and real data.

Deep Dynamical Modeling and Control of Unsteady Fluid Flows

Jeremy Morton, Antony Jameson, Mykel J. Kochenderfer, Freddie Witherden

The design of flow control systems remains a challenge due to the nonlinear nature of the equations that govern fluid flow. However, recent advances in computational fluid dynamics (CFD) have enabled the simulation of complex fluid flows with high accuracy, opening the possibility of using learning-based approaches to facilitate controller design. We present a method for learning the forced and unforced dynamics of airflow over a cylinder directly from CFD data. The proposed approach, grounded in Koopman theory, is shown to produce stable dynamical models that can predict the time evolution of the cylinder system over extended time horizons. Finally, by performing model predictive control with the learned dynamical models, we are able to find a straightforward, interpretable control law for

r suppressing vortex shedding in the wake of the cylinder.

Learning Deep Disentangled Embeddings With the F-Statistic Loss

Karl Ridgeway, Michael C. Mozer

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Disconnected Manifold Learning for Generative Adversarial Networks

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 shortcomings of GANs, its consequences in practice, and the effectiveness of our proposed modifications in alleviating these issues.

Automating Bayesian optimization with Bayesian optimization

Gustavo Malkomes, Roman Garnett

Bayesian optimization is a powerful tool for global optimization of expensive functions. One of its key components is the underlying probabilistic model used for the objective function f . In practice, however, it is often unclear how one should appropriately choose a model, especially when gathering data is expensive. In this work, we introduce a novel automated Bayesian optimization approach that dynamically selects promising models for explaining the observed data using Bayesian Optimization in the model space. Crucially, we account for the uncertainty in the choice of model; our method is capable of using multiple models to represent its current belief about f and subsequently using this information for decision making. We argue, and demonstrate empirically, that our approach automatically finds suitable models for the objective function, which ultimately results in more-efficient optimization.

Leveraged volume sampling for linear regression

Michal Dereziński, Manfred K. K. Warmuth, Daniel J. Hsu

Suppose an $n \times d$ design matrix in a linear regression problem is given, but the response for each point is hidden unless explicitly requested.

The goal is to sample only a small number $k \ll n$ of the responses, and then produce a weight vector whose sum of squares loss over all points is at most $1+\epsilon$ times the minimum.

When k is very small (e.g., $k=d$), jointly sampling diverse subsets of points is crucial. One such method called "volume sampling" has a unique and desirable property that the weight vector it produces is an unbiased

estimate of the optimum. It is therefore natural to ask if this method offers the optimal unbiased estimate in terms of the number of responses k needed to achieve a $1+\epsilon$ loss approximation.

Scalable Robust Matrix Factorization with Nonconvex Loss

Quanming Yao, James Kwok

Requests for name changes in the electronic proceedings will be accepted with no

questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Pipe-SGD: A Decentralized Pipelined SGD Framework for Distributed Deep Net Training

Youjie Li, Mingchao Yu, Songze Li, Salman Avestimehr, Nam Sung Kim, Alexander Schwing

Distributed training of deep nets is an important technique to address some of the present day computing challenges like memory consumption and computational demands. Classical distributed approaches, synchronous or asynchronous, are based on the parameter server architecture, i.e., worker nodes compute gradients which are communicated to the parameter server while updated parameters are returned.

Recently, distributed training with AllReduce operations gained popularity as well. While many of those operations seem appealing, little is reported about wall-clock training time improvements. In this paper, we carefully analyze the AllReduce based setup, propose timing models which include network latency, bandwidth, cluster size and compute time, and demonstrate that a pipelined training with a width of two combines the best of both synchronous and asynchronous training.

Specifically, for a setup consisting of a four-node GPU cluster we show wall-clock time training improvements of up to 5.4x compared to conventional approaches.

Wasserstein Variational Inference

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Marcel A. J. van Gerven, Eric Maris

This paper introduces Wasserstein variational inference, a new form of approximate Bayesian inference based on optimal transport theory. Wasserstein variational inference uses a new family of divergences that includes both f-divergences and the Wasserstein distance as special cases. The gradients of the Wasserstein variational loss are obtained by backpropagating through the Sinkhorn iterations. This technique results in a very stable likelihood-free training method that can be used with implicit distributions and probabilistic programs. Using the Wasserstein variational inference framework, we introduce several new forms of autoencoders and test their robustness and performance against existing variational autoencoding techniques.

Bayesian Control of Large MDPs with Unknown Dynamics in Data-Poor Environments

Mahdi Imani, Seyede Fatemeh Ghoreishi, Ulisses M. Braga-Neto

We propose a Bayesian decision making framework for control of Markov Decision Processes (MDPs) with unknown dynamics and large, possibly continuous, state, action, and parameter spaces in data-poor environments. Most of the existing adaptive controllers for MDPs with unknown dynamics are based on the reinforcement learning framework and rely on large data sets acquired by sustained direct interaction with the system or via a simulator. This is not feasible in many applications, due to ethical, economic, and physical constraints. The proposed framework addresses the data poverty issue by decomposing the problem into an offline planning stage that does not rely on sustained direct interaction with the system or simulator and an online execution stage. In the offline process, parallel Gaussian process temporal difference (GPTD) learning techniques are employed for near-optimal Bayesian approximation of the expected discounted reward over a sample drawn from the prior distribution of unknown parameters. In the online stage, the action with the maximum expected return with respect to the posterior distribution of the parameters is selected. This is achieved by an approximation of the posterior distribution using a Markov Chain Monte Carlo (MCMC) algorithm, followed by constructing multiple Gaussian processes over the parameter space for efficient prediction of the means of the expected return at the MCMC sample. The effectiveness of the proposed framework is demonstrated using a simple dynamical system model with continuous state and action spaces, as well as a more complex model for a metastatic melanoma gene regulatory network observed through noisy synt

hetic gene expression data.

A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem

Sampath Kannan, Jamie H. Morgenstern, Aaron Roth, Bo Waggoner, Zhiwei Steven Wu
Bandit learning is characterized by the tension between long-term exploration and short-term exploitation. However, as has recently been noted, in settings in which the choices of the learning algorithm correspond to important decisions about individual people (such as criminal recidivism prediction, lending, and sequential drug trials), exploration corresponds to explicitly sacrificing the well-being of one individual for the potential future benefit of others. In such settings, one might like to run a "greedy" algorithm, which always makes the optimal decision for the individuals at hand --- but doing this can result in a catastrophic failure to learn. In this paper, we consider the linear contextual bandit problem and revisit the performance of the greedy algorithm.

Lifelong Inverse Reinforcement Learning

Jorge Mendez, Shashank Shivkumar, Eric Eaton

Methods for learning from demonstration (LfD) have shown success in acquiring behavior policies by imitating a user. However, even for a single task, LfD may require numerous demonstrations. For versatile agents that must learn many tasks via demonstration, this process would substantially burden the user if each task were learned in isolation. To address this challenge, we introduce the novel problem of lifelong learning from demonstration, which allows the agent to continually build upon knowledge learned from previously demonstrated tasks to accelerate the learning of new tasks, reducing the amount of demonstrations required. As one solution to this problem, we propose the first lifelong learning approach to inverse reinforcement learning, which learns consecutive tasks via demonstration, continually transferring knowledge between tasks to improve performance.

Recurrent World Models Facilitate Policy Evolution

David Ha, Jürgen Schmidhuber

A generative recurrent neural network is quickly trained in an unsupervised manner to model popular reinforcement learning environments through compressed spatio-temporal representations. The world model's extracted features are fed into compact and simple policies trained by evolution, achieving state of the art results in various environments. We also train our agent entirely inside of an environment generated by its own internal world model, and transfer this policy back into the actual environment. Interactive version of this paper is available at <https://worldmodels.github.io>

Algorithms and Theory for Multiple-Source Adaptation

Judy Hoffman, Mehryar Mohri, Ningshan Zhang

We present a number of novel contributions to the multiple-source adaptation problem. We derive new normalized solutions with strong theoretical guarantees for the cross-entropy loss and other similar losses. We also provide new guarantees that hold in the case where the conditional probabilities for the source domains are distinct. Moreover, we give new algorithms for determining the distribution-weighted combination solution for the cross-entropy loss and other losses. We report the results of a series of experiments with real-world datasets. We find that our algorithm outperforms competing approaches by producing a single robust model that performs well on any target mixture distribution. Altogether, our theory, algorithms, and empirical results provide a full solution for the multiple-source adaptation problem with very practical benefits.

On preserving non-discrimination when combining expert advice

Avrim Blum, Suriya Gunasekar, Thodoris Lykouris, Nati Srebro

We study the interplay between sequential decision making and avoiding discrimination against protected groups, when examples arrive online and do not follow distributional assumptions. We consider the most basic extension of classical online

ne learning: Given a class of predictors that are individually non-discriminatory with respect to a particular metric, how can we combine them to perform as well as the best predictor, while preserving non-discrimination? Surprisingly we show that this task is unachievable for the prevalent notion of "equalized odds" that requires equal false negative rates and equal false positive rates across groups. On the positive side, for another notion of non-discrimination, "equalized error rates", we show that running separate instances of the classical multiplicative weights algorithm for each group achieves this guarantee. Interestingly, even for this notion, we show that algorithms with stronger performance guarantees than multiplicative weights cannot preserve non-discrimination.

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

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 data sets 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.

The Price of Privacy for Low-rank Factorization

Jalaj Upadhyay

In this paper, we study what price one has to pay to release ϵ -differentially private low-rank factorization of a matrix. We consider various settings that are close to the real world applications of low-rank factorization: (i) the manner in which matrices are updated (row by row or in an arbitrary manner), (ii) whether matrices are distributed or not, and (iii) how the output is produced (once at the end of all updates, also known as ϵ -one-shot algorithms or continually). Even though these settings are well studied without privacy, surprisingly, there are no private algorithm for these settings (except when a matrix is updated row by row). We present the first set of differentially private algorithms for all these settings.

Efficient Formal Safety Analysis of Neural Networks

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 sca

le 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.

Inferring Networks From Random Walk-Based Node Similarities

Jeremy Hoskins, Cameron Musco, Christopher Musco, Babis Tsourakakis

Digital presence in the world of online social media entails significant privacy risks. In this work we consider a privacy threat to a social network in which an attacker has access to a subset of random walk-based node similarities, such as effective resistances (i.e., commute times) or personalized PageRank scores. Using these similarities, the attacker seeks to infer as much information as possible about the network, including unknown pairwise node similarities and edges.

Unsupervised Learning of View-invariant Action Representations

Junnan Li, Yongkang Wong, Qi Zhao, Mohan S. Kankanhalli

The recent success in human action recognition with deep learning methods mostly adopt the supervised learning paradigm, which requires significant amount of manually labeled data to achieve good performance. However, label collection is an expensive and time-consuming process. In this work, we propose an unsupervised learning framework, which exploits unlabeled data to learn video representations. Different from previous works in video representation learning, our unsupervised learning task is to predict 3D motion in multiple target views using video representation from a source view. By learning to extrapolate cross-view motions, the representation can capture view-invariant motion dynamics which is discriminative for the action. In addition, we propose a view-adversarial training method to enhance learning of view-invariant features. We demonstrate the effectiveness of the learned representations for action recognition on multiple datasets.

Extracting Relationships by Multi-Domain Matching

Yitong Li, Michael Murias, Geraldine Dawson, David E. 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.

Distributed k-Clustering for Data with Heavy Noise

Shi Li, Xiangyu Guo

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth

ors prior to requesting a name change in the electronic proceedings.

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

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.

Diverse Ensemble Evolution: Curriculum Data-Model Marriage

Tianyi Zhou, Shengjie Wang, Jeff A. Bilmes

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Q-learning with Nearest Neighbors

Devavrat Shah, Qiaomin Xie

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Modular Networks: Learning to Decompose Neural Computation

Louis Kirsch, Julius Kunze, David Barber

Scaling model capacity has been vital in the success of deep learning. For a typical network, necessary compute resources and training time grow dramatically with model size. Conditional computation is a promising way to increase the number of parameters with a relatively small increase in resources. We propose a training algorithm that flexibly chooses neural modules based on the data to be processed. Both the decomposition and modules are learned end-to-end. In contrast to existing approaches, training does not rely on regularization to enforce diversity in module use. We apply modular networks both to image recognition and language modeling tasks, where we achieve superior performance compared to several baselines. Introspection reveals that modules specialize in interpretable contexts.

The Convergence of Sparsified Gradient Methods

Dan Alistarh, Torsten Hoeftler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, Cedric Renggli

Distributed training of massive machine learning models, in particular deep neural networks, via Stochastic Gradient Descent (SGD) is becoming commonplace. Several families of communication-reduction methods, such as quantization, large-batch methods, and gradient sparsification, have been proposed. To date, gradient sparsification methods--where each node sorts gradients by magnitude, and only communicates a subset of the components, accumulating the rest locally--are known to yield some of the largest practical gains. Such methods can reduce the amount of communication per step by up to $\text{\emph{three orders of magnitude}}$, while preserving model accuracy. Yet, this family of methods currently has no theoretical justification.

Deepcode: Feedback Codes via Deep Learning

Hyeji Kim, Yihan Jiang, Sreeram Kannan, Sewoong Oh, Pramod Viswanath

The design of codes for communicating reliably over a statistically well defined

channel is an important endeavor involving deep mathematical research and wide-ranging practical applications. In this work, we present the first family of codes obtained via deep learning, which significantly beats state-of-the-art codes designed over several decades of research. The communication channel under consideration is the Gaussian noise channel with feedback, whose study was initiated by Shannon; feedback is known theoretically to improve reliability of communication, but no practical codes that do so have ever been successfully constructed.

Chain of Reasoning for Visual Question Answering

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.

Hamiltonian Variational Auto-Encoder

Anthony L. Caterini, Arnaud Doucet, Dino Sejdinovic

Variational Auto-Encoders (VAE) have become very popular techniques to perform inference and learning in latent variable models as they allow us to leverage the rich

representational power of neural networks to obtain flexible approximations of the

posterior of latent variables as well as tight evidence lower bounds (ELBO). Combined

with stochastic variational inference, this provides a methodology scaling to

large datasets. However, for this methodology to be practically efficient, it is necessary

to obtain low-variance unbiased estimators of the ELBO and its gradients with

respect to the parameters of interest. While the use of Markov chain Monte Carlo (MCMC) techniques such as Hamiltonian Monte Carlo (HMC) has been previously suggested to achieve this [23, 26], the proposed methods require specifying reverse

kernels which have a large impact on performance. Additionally, the resulting unbiased estimator of the ELBO for most MCMC kernels is typically not amenable to the reparameterization trick. We show here how to optimally select reverse kernels in this setting and, by building upon Hamiltonian Importance Sampling (HIS) [17], we obtain a scheme that provides low-variance unbiased estimators of the ELBO and its gradients using the reparameterization trick. This allows us to develop a Hamiltonian Variational Auto-Encoder (HVAE). This method can be re-interpreted as a target-informed normalizing flow [20] which, within our context,

only requires a few evaluations of the gradient of the sampled likelihood and trivial

Jacobian calculations at each iteration.

Unorganized Malicious Attacks Detection

Ming Pang, Wei Gao, Min Tao, Zhi-Hua Zhou

Recommender systems have attracted much attention during the past decade. Many attack detection algorithms have been developed for better recommendations, mostly focusing on shilling attacks, where an attack organizer produces a large number of user profiles by the same strategy to promote or demote an item. This work considers another different attack style: unorganized malicious attacks, where attackers individually utilize a small number of user profiles to attack different items without organizer. This attack style occurs in many real applications, yet relevant study remains open. We formulate the unorganized malicious attacks detection as a matrix completion problem, and propose the Unorganized Malicious Attacks detection (UMA) algorithm, based on the alternating splitting augmented Lagrangian method. We verify, both theoretically and empirically, the effectiveness of the proposed approach.

Differentially Private k-Means with Constant Multiplicative Error

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.

Multi-value Rule Sets for Interpretable Classification with Feature-Efficient Representations

Tong Wang

We present the Multi-value Rule Set (MRS) for interpretable classification with feature efficient presentations. Compared to rule sets built from single-value rules, MRS adopts a more generalized form of association rules that allows multiple values in a condition. Rules of this form are more concise than classical single-value rules in capturing and describing patterns in data. Our formulation also pursues a higher efficiency of feature utilization, which reduces possible cost in data collection and storage. We propose a Bayesian framework for formulating an MRS model and develop an efficient inference method for learning a maximum a posteriori, incorporating theoretically grounded bounds to iteratively reduce the search space and improve the search efficiency. Experiments on synthetic and real-world data demonstrate that MRS models have significantly smaller complexity and fewer features than baseline models while being competitive in predictive accuracy.

Provable Gaussian Embedding with One Observation

Ming Yu, Zhuoran Yang, Tuo Zhao, Mladen Kolar, Zhaoran Wang

The success of machine learning methods heavily relies on having an appropriate representation for data at hand. Traditionally, machine learning approaches relied on user-defined heuristics to extract features encoding structural information about data. However, recently there has been a surge in approaches that learn how to encode the data automatically in a low dimensional space. Exponential family embedding provides a probabilistic framework for learning low-dimensional representation for various types of high-dimensional data. Though successful in practice, theoretical underpinnings for exponential family embeddings have not been established. In this paper, we study the Gaussian embedding model and develop the first theoretical results for exponential family embedding models. First, we show that, under a mild condition, the embedding structure can be learned from one observation by leveraging the parameter sharing between different contexts even though the data are dependent with each other. Second, we study properties of two algorithms used for learning the embedding structure and establish convergence results for each of them. The first algorithm is based on a convex relaxation, while the other solved the non-convex formulation of the problem directly. Experiments demonstrate the effectiveness of our approach.

Contamination Attacks and Mitigation in Multi-Party Machine Learning

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.

Gradient Sparsification for Communication-Efficient Distributed Optimization

Jianqiao Wangni, Jiale Wang, Ji Liu, Tong Zhang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

The promises and pitfalls of Stochastic Gradient Langevin Dynamics

Nicolas Brosse, Alain Durmus, Eric Moulines

Stochastic Gradient Langevin Dynamics (SGLD) has emerged as a key MCMC algorithm for Bayesian learning from large scale datasets. While SGLD with decreasing step sizes converges weakly to the posterior distribution, the algorithm is often used with a constant step size in practice and has demonstrated spectacular successes in machine learning tasks. The current practice is to set the step size inversely proportional to N where N is the number of training samples. As N becomes large, we show that the SGLD algorithm has an invariant probability measure which significantly departs from the target posterior and behaves like as Stochastic Gradient Descent (SGD). This difference is inherently due to the high variance of the stochastic gradients. Several strategies have been suggested to reduce this effect; among them, SGLD Fixed Point (SGLDFP) uses carefully designed control variates to reduce the variance of the stochastic gradients. We show that SGLDFP gives approximate samples from the posterior distribution, with an accuracy comparable to the Langevin Monte Carlo (LMC) algorithm for a computational cost sublinear in the number of data points. We provide a detailed analysis of the Wasserstein distances between LMC, SGLD, SGLDFP and SGD and explicit expressions of the means and covariance matrices of their invariant distributions. Our findings are supported by limited numerical experiments.

Training Deep Neural Networks with 8-bit Floating Point Numbers

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.

ATOMO: Communication-efficient Learning via Atomic Sparsification

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, Stephen Wright

Distributed model training suffers from communication overheads due to frequent gradient updates transmitted between compute nodes. To mitigate these overheads, several studies propose the use of sparsified stochastic gradients. We argue that these are facets of a general sparsification method that can operate on any possible atomic decomposition. Notable examples include element-wise, singular value, and Fourier decompositions. We present ATOMO, a general framework for atomic sparsification of stochastic gradients. Given a gradient, an atomic decomposition, and a sparsity budget, ATOMO gives a random unbiased sparsification of the atoms minimizing variance. We show that recent methods such as QSGD and TernGrad are special cases of ATOMO, and that sparsifying the singular value decomposition of neural networks gradients, rather than their coordinates, can lead to significantly faster distributed training.

Depth-Limited Solving for Imperfect-Information Games

Noam Brown, Tuomas Sandholm, Brandon Amos

A fundamental challenge in imperfect-information games is that states do not have well-defined values. As a result, depth-limited search algorithms used in single-agent settings and perfect-information games do not apply. This paper introduces a principled way to conduct depth-limited solving in imperfect-information games by allowing the opponent to choose among a number of strategies for the remainder of the game at the depth limit. Each one of these strategies results in a different set of values for leaf nodes. This forces an agent to be robust to the different strategies an opponent may employ. We demonstrate the effectiveness of this approach by building a master-level heads-up no-limit Texas hold'em poker AI that defeats two prior top agents using only a 4-core CPU and 16 GB of memory. Developing such a powerful agent would have previously required a supercomputer.

Causal Inference and Mechanism Clustering of A Mixture of Additive Noise Models

Shoubo Hu, Zhitang Chen, Vahid Partovi Nia, Laiwan CHAN, Yanhui Geng

The inference of the causal relationship between a pair of observed variables is a fundamental problem in science, and most existing approaches are based on one single causal model. In practice, however, observations are often collected from multiple sources with heterogeneous causal models due to certain uncontrollable factors, which renders causal analysis results obtained by a single model skeptical. In this paper, we generalize the Additive Noise Model (ANM) to a mixture model, which consists of a finite number of ANMs, and provide the condition of its causal identifiability. To conduct model estimation, we propose Gaussian Process Partially Observable Model (GPPOM), and incorporate independence enforcement into it to learn latent parameter associated with each observation. Causal inference and clustering according to the underlying generating mechanisms of the mixture model are addressed in this work. Experiments on synthetic and real data demonstrate the effectiveness of our proposed approach.

Adversarial Risk and Robustness: General Definitions and Implications for the Uniform Distribution

Dimitrios Diochnos, Saeed Mahloujifar, Mohammad Mahmoody

We study adversarial perturbations when the instances are uniformly distributed over $\{0,1\}^n$. We study both "inherent" bounds that apply to any problem and any classifier for such a problem as well as bounds that apply to specific problems and specific hypothesis classes.

Analysis of Krylov Subspace Solutions of Regularized Non-Convex Quadratic Problems

Yair Carmon, John C. Duchi

Requests for name changes in the electronic proceedings will be accepted with no

questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Efficient Anomaly Detection via Matrix Sketching

Vatsal Sharan, Parikshit Gopalan, Udi Wieder

We consider the problem of finding anomalies in high-dimensional data using popular PCA based anomaly scores. The naive algorithms for computing these scores explicitly compute the PCA of the covariance matrix which uses space quadratic in the dimensionality of the data. We give the first streaming algorithms that use space that is linear or sublinear in the dimension. We prove general results showing that *any* sketch of a matrix that satisfies a certain operator norm guarantee can be used to approximate these scores. We instantiate these results with powerful matrix sketching techniques such as Frequent Directions and random projections to derive efficient and practical algorithms for these problems, which we validate over real-world data sets. Our main technical contribution is to prove matrix perturbation inequalities for operators arising in the computation of these measures.

Backpropagation with Callbacks: Foundations for Efficient and Expressive Differentiable Programming

Fei Wang, James Decker, Xilun Wu, Gregory Essertel, Tiark Rompf

Training of deep learning models depends on gradient descent and end-to-end differentiation. Under the slogan of differentiable programming, there is an increasing demand for efficient automatic gradient computation for emerging network architectures that incorporate dynamic control flow, especially in NLP.

Constrained Cross-Entropy Method for Safe Reinforcement Learning

Min Wen, Ufuk Topcu

We study a safe reinforcement learning problem in which the constraints are defined as the expected cost over finite-length trajectories. We propose a constrained cross-entropy-based method to solve this problem. The method explicitly tracks its performance with respect to constraint satisfaction and thus is well-suited for safety-critical applications. We show that the asymptotic behavior of the proposed algorithm can be almost-surely described by that of an ordinary differential equation. Then we give sufficient conditions on the properties of this differential equation to guarantee the convergence of the proposed algorithm. At last, we show with simulation experiments that the proposed algorithm can effectively learn feasible policies without assumptions on the feasibility of initial policies, even with non-Markovian objective functions and constraint functions.

Graphical model inference: Sequential Monte Carlo meets deterministic approximations

Fredrik Lindsten, Jouni Helske, Matti Vihola

Approximate inference in probabilistic graphical models (PGMs) can be grouped into deterministic methods and Monte-Carlo-based methods. The former can often provide accurate and rapid inferences, but are typically associated with biases that are hard to quantify. The latter enjoy asymptotic consistency, but can suffer from high computational costs. In this paper we present a way of bridging the gap between deterministic and stochastic inference. Specifically, we suggest an efficient sequential Monte Carlo (SMC) algorithm for PGMs which can leverage the output from deterministic inference methods. While generally applicable, we show explicitly how this can be done with loopy belief propagation, expectation propagation, and Laplace approximations. The resulting algorithm can be viewed as a post-correction of the biases associated with these methods and, indeed, numerical results show clear improvements over the baseline deterministic methods as well as over "plain" SMC.

Playing hard exploration games by watching YouTube

Yusuf Aytaç, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, Nando de Freitas

as

Deep reinforcement learning methods traditionally struggle with tasks where environment rewards are particularly sparse. One successful method of guiding exploration in these domains is to imitate trajectories provided by a human demonstrator. However, these demonstrations are typically collected under artificial conditions, i.e. with access to the agent's exact environment setup and the demonstrator's action and reward trajectories. Here we propose a method that overcomes these limitations in two stages. First, we learn to map unaligned videos from multiple sources to a common representation using self-supervised objectives constructed over both time and modality (i.e. vision and sound). Second, we embed a single YouTube video in this representation to learn a reward function that encourages an agent to imitate human gameplay. This method of one-shot imitation allows our agent to convincingly exceed human-level performance on the infamously hard exploration games Montezuma's Revenge, Pitfall! and Private Eye for the first time, even if the agent is not presented with any environment rewards.

Improved Algorithms for Collaborative PAC Learning

Huy Nguyen, Lydia Zakyntinou

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Scaling provable adversarial defenses

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

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Understanding Batch Normalization

Nils Bjorck, Carla P. Gomes, Bart Selman, Kilian Q. Weinberger

Batch normalization (BN) is a technique to normalize activations in intermediate layers of deep neural networks. Its tendency to improve accuracy and speed up training have established BN as a favorite technique in deep learning. Yet, despite its enormous success, there remains little consensus on the exact reason and mechanism behind these improvements. In this paper we take a step towards a better understanding of BN, following an empirical approach. We conduct several experiments, and show that BN primarily enables training with larger learning rates, which is the cause for faster convergence and better generalization. For networks without BN we demonstrate how large gradient updates can result in diverging loss and activations growing uncontrollably with network depth, which limits possible learning rates. BN avoids this problem by constantly correcting activations to be zero-mean and of unit standard deviation, which enables larger gradient steps, yields faster convergence and may help bypass sharp local minima. We further show various ways in which gradients and activations of deep unnormalized networks are ill-behaved. We contrast our results against recent findings in random matrix theory, shedding new light on classical initialization schemes and their consequences.

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

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 precisely 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.

Learning from discriminative feature feedback

Sanjoy Dasgupta, Akansha Dey, Nicholas Roberts, Sivan Sabato

We consider the problem of learning a multi-class classifier from labels as well as simple explanations that we call "discriminative features". We show that such explanations can be provided whenever the target concept is a decision tree, or more generally belongs to a particular subclass of DNF formulas. We present an efficient online algorithm for learning from such feedback and we give tight bounds on the number of mistakes made during the learning process. These bounds depend only on the size of the target concept and not on the overall number of available features, which could be infinite. We also demonstrate the learning procedure experimentally.

Zeroth-order (Non)-Convex Stochastic Optimization via Conditional Gradient and Gradient Updates

Krishnakumar Balasubramanian, Saeed Ghadimi

In this paper, we propose and analyze zeroth-order stochastic approximation algorithms for nonconvex and convex optimization. Specifically, we propose generalizations of the conditional gradient algorithm achieving rates similar to the standard stochastic gradient algorithm using only zeroth-order information. Furthermore, under a structural sparsity assumption, we first illustrate an implicit regularization phenomenon where the standard stochastic gradient algorithm with zeroth-order information adapts to the sparsity of the problem at hand by just varying the step-size. Next, we propose a truncated stochastic gradient algorithm with zeroth-order information, whose rate of convergence depends only poly-logarithmically on the dimensionality.

Coordinate Descent with Bandit Sampling

Farnood Salehi, Patrick Thiran, Elisa Celis

Coordinate descent methods minimize a cost function by updating a single decision variable (corresponding to one coordinate) at a time. Ideally, we would update the decision variable that yields the largest marginal decrease in the cost function. However, finding this coordinate would require checking all of them, which is not computationally practical. Therefore, we propose a new adaptive method for coordinate descent. First, we define a lower bound on the decrease of the cost function when a coordinate is updated and, instead of calculating this lower bound for all coordinates, we use a multi-armed bandit algorithm to learn which coordinates result in the largest marginal decrease and simultaneously perform coordinate descent. We show that our approach improves the convergence of the coordinate methods both theoretically and experimentally.

PAC-Bayes bounds for stable algorithms with instance-dependent priors

Omar Rivasplata, Emilio Parrado-Hernandez, John S. Shawe-Taylor, Shiliang Sun, Csaba Szepesvari

PAC-Bayes bounds have been proposed to get risk estimates based on a training sample. In this paper the PAC-Bayes approach is combined with stability of the hypothesis learned by a Hilbert space valued algorithm. The PAC-Bayes setting is used with a Gaussian prior centered at the expected output. Thus a novelty of our paper is using priors defined in terms of the data-generating distribution. Our main result estimates the risk of the randomized algorithm in terms of the hypothesis stability coefficients. We also provide a new bound for the SVM classifier, which is compared to other known bounds experimentally. Ours appears to be the first uniform hypothesis stability-based bound that evaluates to non-trivial values.

lues.

DropMax: Adaptive Variational Softmax

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.

Multi-Layered Gradient Boosting Decision Trees

Ji Feng, Yang Yu, Zhi-Hua Zhou

Multi-layered distributed representation is believed to be the key ingredient of deep neural networks especially in cognitive tasks like computer vision. While non-differentiable models such as gradient boosting decision trees (GBDTs) are still the dominant methods for modeling discrete or tabular data, they are hard to incorporate with such representation learning ability. In this work, we propose the multi-layered GBDT forest (mGBDTs), with an explicit emphasis on exploring the ability to learn hierarchical distributed representations by stacking several layers of regression GBDTs as its building block. The model can be jointly trained by a variant of target propagation across layers, without the need to derive backpropagation nor differentiability. Experiments confirmed the effectiveness of the model in terms of performance and representation learning ability.

Rest-Katyusha: Exploiting the Solution's Structure via Scheduled Restart Schemes

Junqi Tang, Mohammad Golbabaee, Francis Bach, Mike E. Davies

We propose a structure-adaptive variant of the state-of-the-art stochastic variance-reduced gradient algorithm Katyusha for regularized empirical risk minimization. The proposed method is able to exploit the intrinsic low-dimensional structure of the solution, such as sparsity or low rank which is enforced by a non-smooth regularization, to achieve even faster convergence rate. This provable algorithmic improvement is done by restarting the Katyusha algorithm according to restricted strong-convexity constants. We demonstrate the effectiveness of our approach via numerical experiments.

Automatic Program Synthesis of Long Programs with a Learned Garbage Collector

Amit Zohar, Lior Wolf

We consider the problem of generating automatic code given sample input-output pairs. We train a neural network to map from the current state and the outputs to the program's next statement. The neural network optimizes multiple tasks concurrently: the next operation out of a set of high level commands, the operands of the next statement, and which variables can be dropped from memory. Using our method we are able to create programs that are more than twice as long as existing state-of-the-art solutions, while improving the success rate for comparable lengths, and cutting the run-time by two orders of magnitude. Our code, including an implementation of various literature baselines, is publicly available at <http://github.com/amitz25/PCCoder>

Quantifying Learning Guarantees for Convex but Inconsistent Surrogates

Kirill Struminsky, Simon Lacoste-Julien, Anton Osokin

We study consistency properties of machine learning methods based on minimizing convex surrogates. We extend the recent framework of Osokin et al. (2017) for th

a quantitative analysis of consistency properties to the case of inconsistent surrogates. Our key technical contribution consists in a new lower bound on the calibration function for the quadratic surrogate, which is non-trivial (not always zero) for inconsistent cases. The new bound allows to quantify the level of inconsistency of the setting and shows how learning with inconsistent surrogates can have guarantees on sample complexity and optimization difficulty. We apply our theory to two concrete cases: multi-class classification with the tree-structured loss and ranking with the mean average precision loss. The results show the approximation-computation trade-offs caused by inconsistent surrogates and their potential benefits.

Unsupervised Text Style Transfer using Language Models as Discriminators

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P. 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.

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

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.

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris M. Mooij

An important goal common to domain adaptation and causal inference is to make accurate predictions when the distributions for the source (or training) domain(s) and target (or test) domain(s) differ. In many cases, these different distributions can be modeled as different contexts of a single underlying system, in which each distribution corresponds to a different perturbation of the system, or in causal terms, an intervention. We focus on a class of such causal domain adapta

tion problems, where data for one or more source domains are given, and the task is to predict the distribution of a certain target variable from measurements of other variables in one or more target domains. We propose an approach for solving these problems that exploits causal inference and does not rely on prior knowledge of the causal graph, the type of interventions or the intervention targets. We demonstrate our approach by evaluating a possible implementation on simulated and real world data.

Confounding-Robust Policy Improvement

Nathan Kallus, Angela Zhou

We study the problem of learning personalized decision policies from observational data while accounting for possible unobserved confounding in the data-generating process. Unlike previous approaches that assume unconfoundedness, i.e., no unobserved confounders affected both treatment assignment and outcomes, we calibrate policy learning for realistic violations of this unverifiable assumption with uncertainty sets motivated by sensitivity analysis in causal inference. Our framework for confounding-robust policy improvement optimizes the minimax regret of a candidate policy against a baseline or reference "status quo" policy, over an uncertainty set around nominal propensity weights. We prove that if the uncertainty set is well-specified, robust policy learning can do no worse than the baseline, and only improve if the data supports it. We characterize the adversarial subproblem and use efficient algorithmic solutions to optimize over parametrized spaces of decision policies such as logistic treatment assignment. We assess our methods on synthetic data and a large clinical trial, demonstrating that confounded selection can hinder policy learning and lead to unwarranted harm, while our robust approach guarantees safety and focuses on well-evidenced improvement.

Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes

Ronan Fruit, Matteo Pirodda, Alessandro Lazaric

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Mesh-TensorFlow: Deep Learning for Supercomputers

Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koo, Anirudh Iyer, Peter Hawkins, HyukJoong 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>

Foreground Clustering for Joint Segmentation and Localization in Videos and Images

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.

Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences

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.

The Description Length of Deep Learning models

Léonard Blier, Yann Ollivier

Deep learning models often have more parameters than observations, and still perform well. This is sometimes described as a paradox. In this work, we show experimentally that despite their huge number of parameters, deep neural networks can compress the data losslessly even when taking the cost of encoding the parameters into account. Such a compression viewpoint originally motivated the use of variational methods in neural networks. However, we show that these variational methods provide surprisingly poor compression bounds, despite being explicitly built to minimize such bounds. This might explain the relatively poor practical performance of variational methods in deep learning. Better encoding methods, imported from the Minimum Description Length (MDL) toolbox, yield much better compression values on deep networks.

Semi-crowdsourced Clustering with Deep Generative Models

Yucen Luo, TIAN TIAN, Jiaxin Shi, Jun Zhu, Bo Zhang

We consider the semi-supervised clustering problem where crowdsourcing provides noisy information about the pairwise comparisons on a small subset of data, i.e., whether a sample pair is in the same cluster. We propose a new approach that includes a deep generative model (DGM) to characterize low-level features of the data, and a statistical relational model for noisy pairwise annotations on its subset. The two parts share the latent variables. To make the model automatically trade-off between its complexity and fitting data, we also develop its fully Ba

yesian variant. The challenge of inference is addressed by fast (natural-gradient) stochastic variational inference algorithms, where we effectively combine variational message passing for the relational part and amortized learning of the DGM under a unified framework. Empirical results on synthetic and real-world data sets show that our model outperforms previous crowdsourced clustering methods.

Scalable Laplacian K-modes

Imtiaz Ziko, Eric Granger, Ismail Ben Ayed

We advocate Laplacian K-modes for joint clustering and density mode finding, and propose a concave-convex relaxation of the problem, which yields a parallel algorithm that scales up to large datasets and high dimensions. We optimize a tight

bound (auxiliary function) of our relaxation, which, at each iteration, amounts to

computing an independent update for each cluster-assignment variable, with guaranteed convergence. Therefore, our bound optimizer can be trivially distributed for large-scale data sets. Furthermore, we show that the density modes can be obtained as byproducts of the assignment variables via simple maximum-value operations whose additional computational cost is linear in the number of data points. Our formulation does not need storing a full affinity matrix and computing

its eigenvalue decomposition, neither does it perform expensive projection steps and Lagrangian-dual inner iterates for the simplex constraints of each point. Furthermore,

unlike mean-shift, our density-mode estimation does not require inner-loop gradient-ascent iterates. It has a complexity independent of feature-space dimension, yields modes that are valid data points in the input set and is applicable

to discrete domains as well as arbitrary kernels. We report comprehensive experiments over various data sets, which show that our algorithm yields very competitive performances in term of optimization quality (i.e., the value of the discrete-variable objective at convergence) and clustering accuracy.

Early Stopping for Nonparametric Testing

Meimei Liu, Guang Cheng

Early stopping of iterative algorithms is an algorithmic regularization method to avoid over-fitting in estimation and classification. In this paper, we show that early stopping can also be applied to obtain the minimax optimal testing in a general non-parametric setup. Specifically, a Wald-type test statistic is obtained based on an iterated estimate produced by functional gradient descent algorithms in a reproducing kernel Hilbert space. A notable contribution is to establish a ``sharp'' stopping rule: when the number of iterations achieves an optimal order, testing optimality is achievable; otherwise, testing optimality becomes impossible. As a by-product, a similar sharpness result is also derived for minimax optimal estimation under early stopping. All obtained results hold for various kernel classes, including Sobolev smoothness classes and Gaussian kernel classes.

Non-Adversarial Mapping with VAEs

Yedid Hoshen

The study of cross-domain mapping without supervision has recently attracted much attention. Much of the recent progress was enabled by the use of adversarial training as well as cycle constraints. The practical difficulty of adversarial training motivates research into non-adversarial methods. In a recent paper, it was shown that cross-domain mapping is possible without the use of cycles or GANs.

Although promising, this approach suffers from several drawbacks including costly inference and an optimization variable for every training example preventing the method from using large training sets. We present an alternative approach which is able to achieve non-adversarial mapping using a novel form of Variational Auto-Encoder. Our method is much faster at inference time, is able to leverage

large datasets and has a simple interpretation.

Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models

Kurtland Chua, Roberto Calandra, Rowan McAllister, Sergey Levine

Model-based reinforcement learning (RL) algorithms can attain excellent sample efficiency, but often lag behind the best model-free algorithms in terms of asymptotic performance. This is especially true with high-capacity parametric function approximators, such as deep networks. In this paper, we study how to bridge this gap, by employing uncertainty-aware dynamics models. We propose a new algorithm called probabilistic ensembles with trajectory sampling (PETS) that combines uncertainty-aware deep network dynamics models with sampling-based uncertainty propagation. Our comparison to state-of-the-art model-based and model-free deep RL algorithms shows that our approach matches the asymptotic performance of model-free algorithms on several challenging benchmark tasks, while requiring significantly fewer samples (e.g. 8 and 125 times fewer samples than Soft Actor Critic and Proximal Policy Optimization respectively on the half-cheetah task).

The challenge of realistic music generation: modelling raw audio at scale

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.

Approximation algorithms for stochastic clustering

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.

Inexact trust-region algorithms on Riemannian manifolds

HiroYuki Kasai, Bamdev Mishra

We consider an inexact variant of the popular Riemannian trust-region algorithm for structured big-data minimization problems. The proposed algorithm approximates the gradient and the Hessian in addition to the solution of a trust-region sub-problem. Addressing large-scale finite-sum problems, we specifically propose sub-sampled algorithms with a fixed bound on sub-sampled Hessian and gradient sizes, where the gradient and Hessian are computed by a random sampling technique. Numerical evaluations demonstrate that the proposed algorithms outperform state-of-the-art Riemannian deterministic and stochastic gradient algorithms across different applications.

Towards Robust Interpretability with Self-Explaining Neural Networks

David Alvarez Melis, Tommi Jaakkola

Most recent work on interpretability of complex machine learning models has focu

sed 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.

Predictive Uncertainty Estimation via Prior Networks

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 *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 *model uncertainty* or as *data uncertainty*. This work proposes a new framework for modeling predictive uncertainty called Prior Networks (PNs) which explicitly models *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.

Graph Oracle Models, Lower Bounds, and Gaps for Parallel Stochastic Optimization

Blake E. Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, Nati Srebro

We suggest a general oracle-based framework that captures parallel stochastic optimization in different parallelization settings described by a dependency graph, and derive generic lower bounds in terms of this graph. We then use the framework and derive lower bounds to study several specific parallel optimization settings, including delayed updates and parallel processing with intermittent communication. We highlight gaps between lower and upper bounds on the oracle complexity, and cases where the "natural" algorithms are not known to be optimal.

An Off-policy Policy Gradient Theorem Using Emphatic Weightings

Ehsan Imani, Eric Graves, Martha White

Policy gradient methods are widely used for control in reinforcement learning, particularly for the continuous action setting. There have been a host of theoretically sound algorithms proposed for the on-policy setting, due to the existence of the policy gradient theorem which provides a simplified form for the gradient. In off-policy learning, however, where the behaviour policy is not necessarily attempting to learn and follow the optimal policy for the given task, the existence of such a theorem has been elusive. In this work, we solve this open problem by providing the first off-policy policy gradient theorem. The key to the derivation is the use of emphatic weightings. We develop a new actor-critic algorithm--called Actor Critic with Emphatic weightings (ACE)--that approximates the simplified gradients provided by the theorem. We demonstrate in a simple counterexample that previous off-policy policy gradient methods--particularly OffPAC and DPG--converge to the wrong solution whereas ACE finds the optimal solution.

Multiple-Step Greedy Policies in Approximate and Online Reinforcement Learning
Yonathan Efroni, Gal Dalal, Bruno Scherrer, Shie Mannor

Multiple-step lookahead policies have demonstrated high empirical competence in Reinforcement Learning, via the use of Monte Carlo Tree Search or Model Predictive Control. In a recent work (Efroni et al., 2018), multiple-step greedy policies and their use in vanilla Policy Iteration algorithms were proposed and analyzed. In this work, we study multiple-step greedy algorithms in more practical setups. We begin by highlighting a counter-intuitive difficulty, arising with soft-policy updates: even in the absence of approximations, and contrary to the 1-step-greedy case, monotonic policy improvement is not guaranteed unless the update stepsize is sufficiently large. Taking particular care about this difficulty, we formulate and analyze online and approximate algorithms that use such a multi-step greedy operator.

Scaling the Poisson GLM to massive neural datasets through polynomial approximations

David Zoltowski, Jonathan W. Pillow

Recent advances in recording technologies have allowed neuroscientists to record simultaneous spiking activity from hundreds to thousands of neurons in multiple brain regions. Such large-scale recordings pose a major challenge to existing statistical methods for neural data analysis. Here we develop highly scalable approximate inference methods for Poisson generalized linear models (GLMs) that require only a single pass over the data. Our approach relies on a recently proposed method for obtaining approximate sufficient statistics for GLMs using polynomial approximations [Huggins et al., 2017], which we adapt to the Poisson GLM setting. We focus on inference using quadratic approximations to nonlinear terms in the Poisson GLM log-likelihood with Gaussian priors, for which we derive closed-form solutions to the approximate maximum likelihood and MAP estimates, posterior distribution, and marginal likelihood. We introduce an adaptive procedure to select the polynomial approximation interval and show that the resulting method allows for efficient and accurate inference and regularization of high-dimensional parameters. We use the quadratic estimator to fit a fully-coupled Poisson GLM to spike train data recorded from 831 neurons across five regions of the mouse brain for a duration of 41 minutes, binned at 1 ms resolution. Across all neurons, this model is fit to over 2 billion spike count bins and identifies fine-timescale statistical dependencies between neurons within and across cortical and subcortical areas.

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

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 spatio-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.

Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, Carlo Ciliberto

Applications of optimal transport have recently gained remarkable attention as a result of the computational advantages of entropic regularization. However, in most situations the Sinkhorn approximation to the Wasserstein distance is replaced by a regularized version that is less accurate but easy to differentiate. In this work we characterize the differential properties of the original Sinkhorn approximation, proving that it enjoys the same smoothness as its regularized version and we explicitly provide an efficient algorithm to compute its gradient. We show that this result benefits both theory and applications: on one hand, high order smoothness confers statistical guarantees to learning with Wasserstein approximations. On the other hand, the gradient formula allows to efficiently solve learning and optimization problems in practice. Promising preliminary experiments complement our analysis.

The Cluster Description Problem - Complexity Results, Formulations and Approximations

Ian Davidson, Antoine Gourru, S Ravi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Global Non-convex Optimization with Discretized Diffusions

Murat A. Erdogdu, Lester Mackey, Ohad Shamir

An Euler discretization of the Langevin diffusion is known to converge to the global minimizers of certain convex and non-convex optimization problems. We show that this property holds for any suitably smooth diffusion and that different diffusions are suitable for optimizing different classes of convex and non-convex functions. This allows us to design diffusions suitable for globally optimizing convex and non-convex functions not covered by the existing Langevin theory. Our non-asymptotic analysis delivers computable optimization and integration error bounds based on easily accessed properties of the objective and chosen diffusion. Central to our approach are new explicit Stein factor bounds on the solutions of Poisson equations. We complement these results with improved optimization guarantees for targets other than the standard Gibbs measure.

Contextual Pricing for Lipschitz Buyers

Jieming Mao, Renato Leme, Jon Schneider

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Processing of missing data by neural networks

Marek Młocieja, Łukasz Struski, Jacek Tabor, Bartosz Zieliński, Przemysław Spurek

We propose a general, theoretically justified mechanism for processing missing data by neural networks. Our idea is to replace typical neuron's response in the first hidden layer by its expected value. This approach can be applied for various types of networks at minimal cost in their modification. Moreover, in contrast to recent approaches, it does not require complete data for training. Experimental results performed on different types of architectures show that our method gives better results than typical imputation strategies and other methods dedicated

ted for incomplete data.

Invariant Representations without Adversarial Training

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, Greg Ver Steeg

Representations of data that are invariant to changes in specified factors are useful for a wide range of problems: removing potential biases in prediction problems, controlling the effects of covariates, and disentangling meaningful factors of variation. Unfortunately, learning representations that exhibit invariance to arbitrary nuisance factors yet remain useful for other tasks is challenging. Existing approaches cast the trade-off between task performance and invariance in an adversarial way, using an iterative minimax optimization. We show that adversarial training is unnecessary and sometimes counter-productive; we instead cast invariant representation learning as a single information-theoretic objective that can be directly optimized. We demonstrate that this approach matches or exceeds performance of state-of-the-art adversarial approaches for learning fair representations and for generative modeling with controllable transformations.

Inference in Deep Gaussian Processes using Stochastic Gradient Hamiltonian Monte Carlo

Marton Havasi, José Miguel Hernández-Lobato, Juan José Murillo-Fuentes

Deep Gaussian Processes (DGPs) are hierarchical generalizations of Gaussian Processes that combine well calibrated uncertainty estimates with the high flexibility of multilayer models. One of the biggest challenges with these models is that exact inference is intractable. The current state-of-the-art inference method, Variational Inference (VI), employs a Gaussian approximation to the posterior distribution. This can be a potentially poor unimodal approximation of the generally multimodal posterior. In this work, we provide evidence for the non-Gaussian nature of the posterior and we apply the Stochastic Gradient Hamiltonian Monte Carlo method to generate samples. To efficiently optimize the hyperparameters, we introduce the Moving Window MCEM algorithm. This results in significantly better predictions at a lower computational cost than its VI counterpart. Thus our method establishes a new state-of-the-art for inference in DGPs.

Regret bounds for meta Bayesian optimization with an unknown Gaussian process prior

Zi Wang, Beomjoon Kim, Leslie Pack Kaelbling

Bayesian optimization usually assumes that a Bayesian prior is given. However, the strong theoretical guarantees in Bayesian optimization are often regrettably compromised in practice because of unknown parameters in the prior. In this paper, we adopt a variant of empirical Bayes and show that, by estimating the Gaussian process prior from offline data sampled from the same prior and constructing unbiased estimators of the posterior, variants of both GP-UCB and $\text{probability of improvement}$ achieve a near-zero regret bound, which decreases to a constant proportional to the observational noise as the number of offline data and the number of online evaluations increase. Empirically, we have verified our approach on challenging simulated robotic problems featuring task and motion planning.

Fully Neural Network Based Speech Recognition on Mobile and Embedded Devices

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 t

he 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.

Large Margin Deep Networks for Classification

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, Samy Bengio

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Collaborative Learning for Deep Neural Networks

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.

Multi-Task Learning as Multi-Objective Optimization

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.

Learning to Exploit Stability for 3D Scene Parsing

Yilun Du, Zhijian Liu, Hector Besevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, Jiajun Wu

Human scene understanding uses a variety of visual and non-visual cues to perform

m inference on object types, poses, and relations. Physics is a rich and universal cue which we exploit to enhance scene understanding. We integrate the physical cue of stability into the learning process using a REINFORCE approach coupled to a physics engine, and apply this to the problem of producing the 3D bounding boxes and poses of objects in a scene. We first show that applying physics supervision to an existing scene understanding model increases performance, produces more stable predictions, and allows training to an equivalent performance level with fewer annotated training examples. We then present a novel architecture for 3D scene parsing named Prim R-CNN, learning to predict bounding boxes as well as their 3D size, translation, and rotation. With physics supervision, Prim R-CNN outperforms existing scene understanding approaches on this problem. Finally, we show that applying physics supervision on unlabeled real images improves real domain transfer of models training on synthetic data.

Direct Runge-Kutta Discretization Achieves Acceleration

Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, Ali Jadbabaie

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Communication Compression for Decentralized Training

Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, Ji Liu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Neural Voice Cloning with a Few Samples

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.

Low-Rank Tucker Decomposition of Large Tensors Using TensorSketch

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.

But How Does It Work in Theory? Linear SVM with Random Features

Yitong Sun, Anna Gilbert, Ambuj Tewari

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

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

Longquan Dai, Liang Tang, Yuan Xie, Jinhui Tang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

A Probabilistic U-Net for Segmentation of Ambiguous Images

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, Olaf Ronneberger

Many real-world vision problems suffer from inherent ambiguities. In clinical applications for example, it might not be clear from a CT scan alone which particular region is cancer tissue. Therefore a group of graders typically produces a set of diverse but plausible segmentations. We consider the task of learning a distribution over segmentations given an input. To this end we propose a generative segmentation model based on a combination of a U-Net with a conditional variational autoencoder that is capable of efficiently producing an unlimited number of plausible hypotheses. We show on a lung abnormalities segmentation task and on a Cityscapes segmentation task that our model reproduces the possible segmentation variants as well as the frequencies with which they occur, doing so significantly better than published approaches. These models could have a high impact in real-world applications, such as being used as clinical decision-making algorithms accounting for multiple plausible semantic segmentation hypotheses to provide possible diagnoses and recommend further actions to resolve the present ambiguities.

Bandit Learning in Concave N-Person Games

Mario Bravo, David Leslie, Panayotis Mertikopoulos

This paper examines the long-run behavior of learning with bandit feedback in non-cooperative concave games. The bandit framework accounts for extremely low-information environments where the agents may not even know they are playing a game; as such, the agents' most sensible choice in this setting would be to employ a no-regret learning algorithm. In general, this does not mean that the players' behavior stabilizes in the long run: no-regret learning may lead to cycles, even with perfect gradient information. However, if a standard monotonicity condition is satisfied, our analysis shows that no-regret learning based on mirror descent with bandit feedback converges to Nash equilibrium with probability 1. We also derive an upper bound for the convergence rate of the process that nearly matches the best attainable rate for single-agent bandit stochastic optimization.

Fast Approximate Natural Gradient Descent in a Kronecker Factored Eigenbasis

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 covariance matrix they are based on becomes gigantic, making them inapplicable in their original form. This has motivated research into both simple diagonal approximations 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 amenable 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.

A convex program for bilinear inversion of sparse vectors

Alireza Aghasi, Ali Ahmed, Paul Hand, Babhru Joshi

We consider the bilinear inverse problem of recovering two vectors, x in \mathbb{R}^L and w in \mathbb{R}^L , from their entrywise product. We consider the case where x and w have known signs and are sparse with respect to known dictionaries of size K and N , respectively. Here, K and N may be larger than, smaller than, or equal to L .

We introduce L1-BranchHull, which is a convex program posed in the natural parameter space and does not require an approximate solution or initialization in order to be stated or solved. We study the case where x and w are S1- and S2-sparse with respect to a random dictionary, with the sparse vectors satisfying an effective sparsity condition, and present a recovery guarantee that depends on the number of measurements as $L > \Omega(S1+S2)(\log(K+N))^2$. Numerical experiments verify that the scaling constant in the theorem is not too large. One application of this problem is the sweep distortion removal task in dielectric imaging, where one of the signals is a nonnegative reflectivity, and the other signal lives in a known subspace, for example that given by dominant wavelet coefficients. We also introduce a variants of L1-BranchHull for the purposes of tolerating noise and outliers, and for the purpose of recovering piecewise constant signals. We provide an ADMM implementation of these variants and show they can extract piecewise constant behavior from real images.

Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks

Yusuke Tsuzuku, Issei Sato, Masashi Sugiyama

High sensitivity of neural networks against malicious perturbations on inputs causes security concerns. To take a steady step towards robust classifiers, we aim to create neural network models provably defended from perturbations. Prior certification work requires strong assumptions on network structures and massive computational costs, and thus the range of their applications was limited. From the relationship between the Lipschitz constants and prediction margins, we present a computationally efficient calculation technique to lower-bound the size of adversarial perturbations that can deceive networks, and that is widely applicable to various complicated networks. Moreover, we propose an efficient training procedure that robustifies networks and significantly improves the provably guarded areas around data points. In experimental evaluations, our method showed its ability to provide a non-trivial guarantee and enhance robustness for even large networks.

Robust Learning of Fixed-Structure Bayesian Networks

Yu Cheng, Ilias Diakonikolas, Daniel Kane, Alistair Stewart

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data
Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, Taco S. 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 \mathbb{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.

Toddler-Inspired Visual Object Learning

Sven Bambach, David Crandall, Linda Smith, Chen Yu

Real-world learning systems have practical limitations on the quality and quantity of the training datasets that they can collect and consider. How should a sys

tem go about choosing a subset of the possible training examples that still allows for learning accurate, generalizable models? To help address this question, we draw inspiration from a highly efficient practical learning system: the human child. Using head-mounted cameras, eye gaze trackers, and a model of foveated vision, we collected first-person (egocentric) images that represents a highly accurate approximation of the "training data" that toddlers' visual systems collect in everyday, naturalistic learning contexts. We used state-of-the-art computer vision learning models (convolutional neural networks) to help characterize the structure of these data, and found that child data produce significantly better object models than egocentric data experienced by adults in exactly the same environment. By using the CNNs as a modeling tool to investigate the properties of the child data that may enable this rapid learning, we found that child data exhibit a unique combination of quality and diversity, with not only many similar large, high-quality object views but also a greater number and diversity of rare views. This novel methodology of analyzing the visual "training data" used by children may not only reveal insights to improve machine learning, but also may suggest new experimental tools to better understand infant learning in developmental psychology.

Reducing Network Agnostophobia

Akshay Raj Dhamija, Manuel Günther, Terrance Boulton

Agnostophobia, the fear of the unknown, can be experienced by deep learning engineers while applying their networks to real-world applications. Unfortunately, network behavior is not well defined for inputs far from a network's training set.

In an uncontrolled environment, networks face many instances that are not of interest to them and have to be rejected in order to avoid a false positive. This problem has previously been tackled by researchers by either a) thresholding softmax, which by construction cannot return "none of the known classes", or b) using an additional background or garbage class. In this paper, we show that both of these approaches help, but are generally insufficient when previously unseen classes are encountered. We also introduce a new evaluation metric that focuses on comparing the performance of multiple approaches in scenarios where such unseen classes or unknowns are encountered. Our major contributions are simple yet effective Entropic Open-Set and Objectosphere losses that train networks using negative samples from some classes. These novel losses are designed to maximize entropy for unknown inputs while increasing separation in deep feature space by modifying magnitudes of known and unknown samples. Experiments on networks trained to classify classes from MNIST and CIFAR-10 show that our novel loss functions are significantly better at dealing with unknown inputs from datasets such as Devanagari, NotMNIST, CIFAR-100 and SVHN.

Deep Functional Dictionaries: Learning Consistent Semantic Structures on 3D Models from Functions

Minhyuk Sung, Hao Su, Ronald Yu, Leonidas J. Guibas

Various 3D semantic attributes such as segmentation masks, geometric features, keypoints, and materials can be encoded as per-point probe functions on 3D geometries. Given a collection of related 3D shapes, we consider how to jointly analyze such probe functions over different shapes, and how to discover common latent structures using a neural network – even in the absence of any correspondence information. Our network is trained on point cloud representations of shape geometry and associated semantic functions on that point cloud. These functions express a shared semantic understanding of the shapes but are not coordinated in any way. For example, in a segmentation task, the functions can be indicator functions of arbitrary sets of shape parts, with the particular combination involved not known to the network. Our network is able to produce a small dictionary of basis functions for each shape, a dictionary whose span includes the semantic functions provided for that shape. Even though our shapes have independent discretizations and no functional correspondences are provided, the network is able to generate latent bases, in a consistent order, that reflect the shared semantic structure among the shapes. We demonstrate the effectiveness of our technique in vari

ous segmentation and keypoint selection applications.

Teaching Inverse Reinforcement Learners via Features and Demonstrations

Luis Haug, Sebastian Tschiatschek, Adish Singla

Learning near-optimal behaviour from an expert's demonstrations typically relies on the assumption that the learner knows the features that the true reward function depends on. In this paper, we study the problem of learning from demonstrations in the setting where this is not the case, i.e., where there is a mismatch between the worldviews of the learner and the expert. We introduce a natural quantity, the teaching risk, which measures the potential suboptimality of policies that look optimal to the learner in this setting. We show that bounds on the teaching risk guarantee that the learner is able to find a near-optimal policy using standard algorithms based on inverse reinforcement learning. Based on these findings, we suggest a teaching scheme in which the expert can decrease the teaching risk by updating the learner's worldview, and thus ultimately enable her to find a near-optimal policy.

Learning to Decompose and Disentangle Representations for Video Prediction

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F. Fei-Fei, Juan Carlos Niebles

Our goal is to predict future video frames given a sequence of input frames. Despite large amounts of video data, this remains a challenging task because of the high-dimensionality of video frames. We address this challenge by proposing the Decompositional Disentangled Predictive Auto-Encoder (DDPAE), a framework that combines structured probabilistic models and deep networks to automatically (i) decompose the high-dimensional video that we aim to predict into components, and (ii) disentangle each component to have low-dimensional temporal dynamics that are easier to predict. Crucially, with an appropriately specified generative model of video frames, our DDPAE is able to learn both the latent decomposition and disentanglement without explicit supervision. For the Moving MNIST dataset, we show that DDPAE is able to recover the underlying components (individual digits) and disentanglement (appearance and location) as we would intuitively do. We further demonstrate that DDPAE can be applied to the Bouncing Balls dataset involving complex interactions between multiple objects to predict the video frame directly from the pixels and recover physical states without explicit supervision.

Maximizing acquisition functions for Bayesian optimization

James Wilson, Frank Hutter, Marc Deisenroth

Bayesian optimization is a sample-efficient approach to global optimization that relies on theoretically motivated value heuristics (acquisition functions) to guide its search process. Fully maximizing acquisition functions produces the Bayes' decision rule, but this ideal is difficult to achieve since these functions are frequently non-trivial to optimize. This statement is especially true when evaluating queries in parallel, where acquisition functions are routinely non-convex, high-dimensional, and intractable. We first show that acquisition functions estimated via Monte Carlo integration are consistently amenable to gradient-based optimization. Subsequently, we identify a common family of acquisition functions, including EI and UCB, whose characteristics not only facilitate but justify use of greedy approaches for their maximization.

Nonparametric Density Estimation under Adversarial Losses

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, Barnabas Póczos

We study minimax convergence rates of nonparametric density estimation under a large class of loss functions called ``adversarial losses'', which, besides classical L^p losses, includes maximum mean discrepancy (MMD), Wasserstein distance, and total variation distance. These losses are closely related to the losses encoded by discriminator networks in generative adversarial networks (GANs). In a general framework, we study how the choice of loss and the assumed smoothness of the underlying density together determine the minimax rate. We also discuss implications for training GANs based on deep ReLU networks, and more general connect

ions to learning implicit generative models in a minimax statistical sense.

Weakly Supervised Dense Event Captioning in Videos

Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, Junzhou Huang
Dense event captioning aims to detect and describe all events of interest contained in a video. Despite the advanced development in this area, existing methods tackle this task by making use of dense temporal annotations, which is dramatically source-consuming. This paper formulates a new problem: weakly supervised dense event captioning, which does not require temporal segment annotations for model training. Our solution is based on the one-to-one correspondence assumption, each caption describes one temporal segment, and each temporal segment has one caption, which holds in current benchmark datasets and most real world cases. We decompose the problem into a pair of dual problems: event captioning and sentence localization and present a cycle system to train our model. Extensive experimental results are provided to demonstrate the ability of our model on both dense event captioning and sentence localization in videos.

Moonshine: Distilling with Cheap Convolutions

Elliot J. Crowley, Gavin Gray, Amos J. 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.

Exploiting Numerical Sparsity for Efficient Learning : Faster Eigenvector Computation and Regression

Neha Gupta, Aaron Sidford

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

The Everlasting Database: Statistical Validity at a Fair Price

Blake E. Woodworth, Vitaly Feldman, Saharon Rosset, Nati Srebro

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning Conditioned Graph Structures for Interpretable Visual Question Answering

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

ain promising results with 66.18% accuracy and demonstrate the interpretability of the proposed method. Code can be found at github.com/aimbrain/vqa-project.

Exponentially Weighted Imitation Learning for Batched Historical Data

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.

Temporal Regularization for Markov Decision Process

Pierre Thodoroff, Audrey Durand, Joelle Pineau, Doina Precup

Several applications of Reinforcement Learning suffer from instability due to high

variance. This is especially prevalent in high dimensional domains. Regularization

is a commonly used technique in machine learning to reduce variance, at the cost of introducing some bias. Most existing regularization techniques focus on spatial

(perceptual) regularization. Yet in reinforcement learning, due to the nature of the

Bellman equation, there is an opportunity to also exploit temporal regularization

based on smoothness in value estimates over trajectories. This paper explores a class of methods for temporal regularization. We formally characterize the bias induced by this technique using Markov chain concepts. We illustrate the various characteristics of temporal regularization via a sequence of simple discrete and continuous MDPs, and show that the technique provides improvement even in high-dimensional Atari games.

Explaining Deep Learning Models -- A Bayesian Non-parametric Approach

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.

Optimization of Smooth Functions with Noisy Observations: Local Minimax Rates

Yining Wang, Sivaraman Balakrishnan, Aarti Singh

We consider the problem of global optimization of an unknown non-convex smooth function with noisy zeroth-order feedback. We propose a local minimax framework to study the fundamental difficulty of optimizing smooth functions with adaptive function evaluations. We show that for functions with fast growth around their global minima, carefully designed optimization algorithms can identify a near global minimizer with many fewer queries than worst-case global minimax theory pred

icts. For the special case of strongly convex and smooth functions, our implied convergence rates match the ones developed for zeroth-order convex optimization problems. On the other hand, we show that in the worst case no algorithm can converge faster than the minimax rate of estimating an unknown functions in linf-norm. Finally, we show that non-adaptive algorithms, although optimal in a global minimax sense, do not attain the optimal local minimax rate.

Measures of distortion for machine learning

Leena Chennuru Vankadara, Ulrike von Luxburg

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Sparse DNNs with Improved Adversarial Robustness

Yiwen Guo, Chao Zhang, Changshui Zhang, Yurong Chen

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

e-SNLI: Natural Language Inference with Natural Language Explanations

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

Synaptic Strength For Convolutional Neural Network

CHEN LIN, Zhao Zhong, Wu Wei, Junjie Yan

Convolutional Neural Networks(CNNs) are both computation and memory intensive which hindered their deployment in mobile devices. Inspired by the relevant concept in neural science literature, we propose Synaptic Pruning: a data-driven method to prune connections between input and output feature maps with a newly proposed class of parameters called Synaptic Strength. Synaptic Strength is designed to capture the importance of a connection based on the amount of information it transports. Experiment results show the effectiveness of our approach. On CIFAR-10, we prune connections for various CNN models with up to 96%, which results in significant size reduction and computation saving. Further evaluation on ImageNet demonstrates that synaptic pruning is able to discover efficient models which is competitive to state-of-the-art compact CNNs such as MobileNet-V2 and NasNet-Mobile. Our contribution is summarized as following: (1) We introduce Synaptic Strength, a new class of parameters for CNNs to indicate the importance of each connections. (2) Our approach can prune various CNNs with high compression without compromising accuracy. (3) Further investigation shows, the proposed Synaptic Strength is a better indicator for kernel pruning compared with the previous approach in both empirical result and theoretical analysis.

Meta-Reinforcement Learning of Structured Exploration Strategies

Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, Sergey Levine

Exploration is a fundamental challenge in reinforcement learning (RL). Many current exploration methods for deep RL use task-agnostic objectives, such as

information gain or bonuses based on state visitation. However, many practical applications of RL involve learning more than a single task, and prior tasks can be used to inform how exploration should be performed in new tasks. In this work, we study how prior tasks can inform an agent about how to explore effectively in new situations. We introduce a novel gradient-based fast adaptation algorithm - model agnostic exploration with structured noise (MAESN) - to learn exploration strategies from prior experience. The prior experience is used both to initialize a policy and to acquire a latent exploration space that can inject structured stochasticity into a policy, producing exploration strategies that are informed by prior knowledge and are more effective than random action-space noise. We show that MAESN is more effective at learning exploration strategies when compared to prior meta-RL methods, RL without learned exploration strategies, and task-agnostic exploration methods. We evaluate our method on a variety of simulated tasks: locomotion with a wheeled robot, locomotion with a quadrupedal walker, and object manipulation.

Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence
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.

A Block Coordinate Ascent Algorithm for Mean-Variance Optimization

Tengyang Xie, Bo Liu, Yangyang Xu, Mohammad Ghavamzadeh, Yinlam Chow, Daoming Lyu, Daesub Yoon

Risk management in dynamic decision problems is a primary concern in many fields, including financial investment, autonomous driving, and healthcare. The mean-variance function is one of the most widely used objective functions in risk management due to its simplicity and interpretability. Existing algorithms for mean-variance optimization are based on multi-time-scale stochastic approximation, whose learning rate schedules are often hard to tune, and have only asymptotic convergence proof. In this paper, we develop a model-free policy search framework for mean-variance optimization with finite-sample error bound analysis (to local optima). Our starting point is a reformulation of the original mean-variance function with its Fenchel dual, from which we propose a stochastic block coordinate ascent policy search algorithm. Both the asymptotic convergence guarantee of the last iteration's solution and the convergence rate of the randomly picked solution are provided, and their applicability is demonstrated on several benchmark domains.

MetaGAN: An Adversarial Approach to Few-Shot Learning

Ruixiang ZHANG, Tong Che, Zoubin Ghahramani, Yoshua Bengio, Yangqiu Song

In this paper, we propose a conceptually simple and general framework called MetaGAN for few-shot learning problems. Most state-of-the-art few-shot classification models can be integrated with MetaGAN in a principled and straightforward way. By introducing an adversarial generator conditioned on tasks, we augment vanilla few-shot classification models with the ability to discriminate between real and fake data. We argue that this GAN-based approach can help few-shot classifiers to learn sharper decision boundary, which could generalize better. We show that with our MetaGAN framework, we can extend supervised few-shot learning models to naturally cope with unsupervised data. Different from previous work in semi-supervised few-shot learning, our algorithms can deal with semi-supervision at both sample-level and task-level. We give theoretical justifications of the strength of MetaGAN, and validate the effectiveness of MetaGAN on challenging few-shot image classification benchmarks.

Efficient High Dimensional Bayesian Optimization with Additivity and Quadrature Fourier Features

Mojmir Mutny, Andreas Krause

We develop an efficient and provably no-regret Bayesian optimization (BO) algorithm for optimization of black-box functions in high dimensions. We assume a generalized additive model with possibly overlapping variable groups. When the groups do not overlap, we are able to provide the first provably no-regret $\text{\emph{poly}}(\log n)$ (in the number of evaluations of the acquisition function) algorithm for solving high dimensional BO. To make the optimization efficient and feasible, we introduce a novel deterministic Fourier Features approximation based on numerical integration with detailed analysis for the squared exponential kernel. The error of this approximation decreases $\text{\emph{exponentially}}$ with the number of features, and allows for a precise approximation of both posterior mean and variance. In addition, the kernel matrix inversion improves in its complexity from cubic to essentially linear in the number of data points measured in basic arithmetic operations.

The Physical Systems Behind Optimization Algorithms

Lin Yang, Raman Arora, Vladimir Braverman, Tuo Zhao

We use differential equations based approaches to provide some $\text{\textit{physical}}$ insights into analyzing the dynamics of popular optimization algorithms in machine learning. In particular, we study gradient descent, proximal gradient descent, coordinate gradient descent, proximal coordinate gradient, and Newton's methods as well as their Nesterov's accelerated variants in a unified framework motivated by a natural connection of optimization algorithms to physical systems. Our analysis is applicable to more general algorithms and optimization problems $\text{\textit{beyond}}$ convexity and strong convexity, e.g. Polyak-Lojasiewicz and error bound conditions (possibly nonconvex).

Unsupervised Learning of Shape and Pose with Differentiable Point Clouds

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.

Unsupervised Attention-guided Image-to-Image Translation

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

Current unsupervised image-to-image translation techniques struggle to focus the 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.

Learning Concave Conditional Likelihood Models for Improved Analysis of Tandem Mass Spectra

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.

Beyond Grids: Learning Graph Representations for Visual Recognition

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.

Approximate Knowledge Compilation by Online Collapsed Importance Sampling

Tal Friedman, Guy Van den Broeck

We introduce collapsed compilation, a novel approximate inference algorithm for discrete probabilistic graphical models. It is a collapsed sampling algorithm that incrementally selects which variable to sample next based on the partial compilation obtained so far. This online collapsing, together with knowledge compilation inference on the remaining variables, naturally exploits local structure and context-specific independence in the distribution. These properties are used implicitly in exact inference, but are difficult to harness for approximate inference. Moreover, by having a partially compiled circuit available during sampling, collapsed compilation has access to a highly effective proposal distribution for importance sampling. Our experimental evaluation shows that collapsed co

mpilation performs well on standard benchmarks. In particular, when the amount of exact inference is equally limited, collapsed compilation is competitive with the state of the art, and outperforms it on several benchmarks.

Layer-Wise Coordination between Encoder and Decoder for Neural Machine Translation

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.

A Lyapunov-based Approach to Safe Reinforcement Learning

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

In many real-world reinforcement learning (RL) problems, besides optimizing the main objective function, an agent must concurrently avoid violating a number of constraints. In particular, besides optimizing performance, it is crucial to guarantee the safety of an agent during training as well as deployment (e.g., a robot should avoid taking actions - exploratory or not - which irreversibly harm its hardware). To incorporate safety in RL, we derive algorithms under the framework of constrained Markov decision processes (CMDPs), an extension of the standard Markov decision processes (MDPs) augmented with constraints on expected cumulative costs. Our approach hinges on a novel Lyapunov method. We define and present a method for constructing Lyapunov functions, which provide an effective way to guarantee the global safety of a behavior policy during training via a set of local linear constraints. Leveraging these theoretical underpinnings, we show how to use the Lyapunov approach to systematically transform dynamic programming (DP) and RL algorithms into their safe counterparts. To illustrate their effectiveness, we evaluate these algorithms in several CMDP planning and decision-making tasks on a safety benchmark domain. Our results show that our proposed method significantly outperforms existing baselines in balancing constraint satisfaction and performance.

Reversible Recurrent Neural Networks

Matthew MacKay, Paul Vicol, Jimmy Ba, Roger B. 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.

Deep Poisson gamma dynamical systems

Dandan Guo, Bo Chen, Hao Zhang, Mingyuan Zhou

We develop deep Poisson-gamma dynamical systems (DPGDS) to model sequentially ob

served multivariate count data, improving previously proposed models by not only mining deep hierarchical latent structure from the data, but also capturing both first-order and long-range temporal dependencies. Using sophisticated but simple-to-implement data augmentation techniques, we derived closed-form Gibbs sampling update equations by first backward and upward propagating auxiliary latent counts, and then forward and downward sampling latent variables. Moreover, we develop stochastic gradient MCMC inference that is scalable to very long multivariate count time series. Experiments on both synthetic and a variety of real-world data demonstrate that the proposed model not only has excellent predictive performance, but also provides highly interpretable multilayer latent structure to represent hierarchical and temporal information propagation.

Regularization Learning Networks: Deep Learning for Tabular Datasets

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 revealing 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>.

Online Learning with an Unknown Fairness Metric

Stephen Gillen, Christopher Jung, Michael Kearns, Aaron Roth

We consider the problem of online learning in the linear contextual bandits setting, but in which there are also strong individual fairness constraints governed by an unknown similarity metric. These constraints demand that we select similar actions or individuals with approximately equal probability, which may be at odds with optimizing reward, thus modeling settings where profit and social policy are in tension. We assume we learn about an unknown Mahalanobis similarity metric from only weak feedback that identifies fairness violations, but does not quantify their extent. This is intended to represent the interventions of a regulator who "knows unfairness when he sees it" but nevertheless cannot enumerate a quantitative fairness metric over individuals. Our main result is an algorithm in the adversarial context setting that has a number of fairness violations that depends only logarithmically on T , while obtaining an optimal $O(\sqrt{T})$ regret bound to the best fair policy.

Completing State Representations using Spectral Learning

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 an

gles 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.

From Stochastic Planning to Marginal MAP

Hao(Jackson) Cui, Radu Marinescu, Roni Kharden

It is well known that the problems of stochastic planning and probabilistic inference are closely related. This paper makes two contributions in this context. The first is to provide an analysis of the recently developed SOGBOFA heuristic planning algorithm that was shown to be effective for problems with large factored state and action spaces. It is shown that SOGBOFA can be seen as a specialized inference algorithm that computes its solutions through a combination of a symbolic variant of belief propagation and gradient ascent. The second contribution is a new solver for Marginal MAP (MMAP) inference. We introduce a new reduction from MMAP to maximum expected utility problems which are suitable for the symbolic computation in SOGBOFA. This yields a novel algebraic gradient-based solver (AGS) for MMAP. An experimental evaluation illustrates the potential of AGS in solving difficult MMAP problems.

Generalizing Graph Matching beyond Quadratic Assignment Model

Tianshu Yu, Junchi Yan, Yilin Wang, Wei Liu, baoxin Li

Graph matching has received persistent attention over decades, which can be formulated as a quadratic assignment problem (QAP). We show that a large family of functions, which we define as Separable Functions, can approximate discrete graph matching in the continuous domain asymptotically by varying the approximation controlling parameters. We also study the properties of global optimality and devise convex/concave-preserving extensions to the widely used Lawler's QAP form. Our theoretical findings show the potential for deriving new algorithms and techniques for graph matching. We deliver solvers based on two specific instances of Separable Functions, and the state-of-the-art performance of our method is verified on popular benchmarks.

On Learning Intrinsic Rewards for Policy Gradient Methods

Zeyu Zheng, Junhyuk Oh, Satinder Singh

In many sequential decision making tasks, it is challenging to design reward functions that help an RL agent efficiently learn behavior that is considered good by the agent designer. A number of different formulations of the reward-design problem, or close variants thereof, have been proposed in the literature. In this paper we build on the Optimal Rewards Framework of Singh et al. that defines the optimal intrinsic reward function as one that when used by an RL agent achieves behavior that optimizes the task-specifying or extrinsic reward function. Previous work in this framework has shown how good intrinsic reward functions can be learned for lookahead search based planning agents. Whether it is possible to learn intrinsic reward functions for learning agents remains an open problem. In this paper we derive a novel algorithm for learning intrinsic rewards for policy-gradient based learning agents. We compare the performance of an augmented agent that uses our algorithm to provide additive intrinsic rewards to an A2C-based policy learner (for Atari games) and a PPO-based policy learner (for Mujoco domains) with a baseline agent that uses the same policy learners but with only extrinsic rewards. Our results show improved performance on most but not all of the domains.

Regularizing by the Variance of the Activations' Sample-Variances

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 acti

variations 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.

Single-Agent Policy Tree Search With Guarantees

Laurent Orseau, Levi Lelis, Tor Lattimore, Theophane Weber

We introduce two novel tree search algorithms that use a policy to guide search. The first algorithm is a best-first enumeration that uses a cost function that allows us to provide an upper bound on the number of nodes to be expanded before reaching a goal state. We show that this best-first algorithm is particularly well suited for ``needle-in-a-haystack'' problems. The second algorithm, which is based on sampling, provides an upper bound on the expected number of nodes to be expanded before reaching a set of goal states. We show that this algorithm is better suited for problems where many paths lead to a goal. We validate these tree search algorithms on 1,000 computer-generated levels of Sokoban, where the policy used to guide search comes from a neural network trained using A3C. Our results show that the policy tree search algorithms we introduce are competitive with a state-of-the-art domain-independent planner that uses heuristic search.

Bias and Generalization in Deep Generative Models: An Empirical Study

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.

Joint Autoregressive and Hierarchical Priors for Learned Image Compression

David Minnen, Johannes Ballé, George D. Toderici

Recent models for learned image compression are based on autoencoders that learn approximately invertible mappings from pixels to a quantized latent representation. The transforms are combined with an entropy model, which is a prior on the latent representation that can be used with standard arithmetic coding algorithms to generate a compressed bitstream. Recently, hierarchical entropy models were introduced as a way to exploit more structure in the latents than previous fully factorized priors, improving compression performance while maintaining end-to-end optimization. Inspired by the success of autoregressive priors in probabilistic generative models, we examine autoregressive, hierarchical, and combined priors as alternatives, weighing their costs and benefits in the context of image compression. While it is well known that autoregressive models can incur a significant computational penalty, we find that in terms of compression performance, autoregressive and hierarchical priors are complementary and can be combined to exploit the probabilistic structure in the latents better than all previous learned models. The combined model yields state-of-the-art rate-distortion performance and generates smaller files than existing methods: 15.8% rate reductions over the baseline hierarchical model and 59.8%, 35%, and 8.4% savings over JPEG, JPEG 2000, and BPG, respectively. To the best of our knowledge, our model is the first learning-based method to outperform the top standard image codec (BPG) on both the PSNR and MS-SSIM distortion metrics.

Link Prediction Based on Graph Neural Networks

Muhan Zhang, Yixin Chen

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

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

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.

A probabilistic population code based on neural samples

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 & Hyvarinen 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.

Generative Probabilistic Novelty Detection with Adversarial Autoencoders

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.

Monte-Carlo Tree Search for Constrained POMDPs

Jongmin Lee, Geon-hyeong Kim, Pascal Poupart, Kee-Eung Kim

Monte-Carlo Tree Search (MCTS) has been successfully applied to very large POMDPs, a standard model for stochastic sequential decision-making problems. However, many real-world problems inherently have multiple goals, where multi-objective formulations are more natural. The constrained POMDP (CPOMDP) is such a model that maximizes the reward while constraining the cost, extending the standard POMDP model. To date, solution methods for CPOMDPs assume an explicit model of the environment, and thus are hardly applicable to large-scale real-world problems. In this paper, we present CC-POMCP (Cost-Constrained POMCP), an online MCTS algorithm for large CPOMDPs that leverages the optimization of LP-induced parameters and only requires a black-box simulator of the environment. In the experiments, we demonstrate that CC-POMCP converges to the optimal stochastic action selection in CPOMDP and pushes the state-of-the-art by being able to scale to very large problems.

Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data

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.

Informative Features for Model Comparison

Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, Arthur Gretton

Given two candidate models, and a set of target observations, we address the problem of measuring the relative goodness of fit of the two models. We propose two new statistical tests which are nonparametric, computationally efficient (runtime complexity is linear in the sample size), and interpretable. As a unique advantage, our tests can produce a set of examples (informative features) indicating the regions in the data domain where one model fits significantly better than the other. In a real-world problem of comparing GAN models, the test power of our new test matches that of the state-of-the-art test of relative goodness of fit, while being one order of magnitude faster.

Discrimination-aware Channel Pruning for Deep Neural Networks

Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, Jinhui Zhu

Channel pruning is one of the predominant approaches for deep model compression. Existing pruning methods either train from scratch with sparsity constraints on channels, or minimize the reconstruction error between the pre-trained feature maps and the compressed ones. Both strategies suffer from some limitations: the former kind is computationally expensive and difficult to converge, whilst the latter kind optimizes the reconstruction error but ignores the discriminative po

wer of channels. To overcome these drawbacks, we investigate a simple-yet-effective method, called discrimination-aware channel pruning, to choose those channels that really contribute to discriminative power. To this end, we introduce additional losses into the network to increase the discriminative power of intermediate layers and then select the most discriminative channels for each layer by considering the additional loss and the reconstruction error. Last, we propose a greedy algorithm to conduct channel selection and parameter optimization in an iterative way. Extensive experiments demonstrate the effectiveness of our method. For example, on ILSVRC-12, our pruned ResNet-50 with 30% reduction of channels even outperforms the original model by 0.39% in top-1 accuracy.

Reinforcement Learning of Theorem Proving

Cezary Kaliszyk, Josef Urban, Henryk Michalewski, Miroslav Olšák

We introduce a theorem proving algorithm that uses practically no domain heuristics for guiding its connection-style proof search. Instead, it runs many Monte-Carlo simulations guided by reinforcement learning from previous proof attempts. We produce several versions of the prover, parameterized by different learning and guiding algorithms. The strongest version of the system is trained on a large corpus of mathematical problems and evaluated on previously unseen problems. The trained system solves within the same number of inferences over 40% more problems than a baseline prover, which is an unusually high improvement in this hard AI domain. To our knowledge this is the first time reinforcement learning has been convincingly applied to solving general mathematical problems on a large scale.

On Fast Leverage Score Sampling and Optimal Learning

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, Lorenzo Rosasco

Leverage score sampling provides an appealing way to perform approximate computations for large matrices. Indeed, it allows to derive faithful approximations with a complexity adapted to the problem at hand. Yet, performing leverage score sampling is a challenge in its own right requiring further approximations. In this paper, we study the problem of leverage score sampling for positive definite matrices defined by a kernel. Our contribution is twofold. First we provide a novel algorithm for leverage score sampling and second, we exploit the proposed method in statistical learning by deriving a novel solver for kernel ridge regression. Our main technical contribution is showing that the proposed algorithms are currently the most efficient and accurate for these problems.

Robustness of conditional GANs to noisy labels

Kiran K. 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.

Removing Hidden Confounding by Experimental Grounding

Nathan Kallus, Aahlad Manas Puli, Uri Shalit

Observational data is increasingly used as a means for making individual-level causal predictions and intervention recommendations. The foremost challenge of causal inference from observational data is hidden confounding, whose presence can not be tested in data and can invalidate any causal conclusion. Experimental data does not suffer from confounding but is usually limited in both scope and scale. We introduce a novel method of using limited experimental data to correct the hidden confounding in causal effect models trained on larger observational data, even if the observational data does not fully overlap with the experimental data. Our method makes strictly weaker assumptions than existing approaches, and we prove conditions under which it yields a consistent estimator. We demonstrate our method's efficacy using real-world data from a large educational experiment.

Legendre Decomposition for Tensors

Mahito Sugiyama, Hiroyuki Nakahara, Koji Tsuda

We present a novel nonnegative tensor decomposition method, called Legendre decomposition, which factorizes an input tensor into a multiplicative combination of parameters. Thanks to the well-developed theory of information geometry, the reconstructed tensor is unique and always minimizes the KL divergence from an input tensor. We empirically show that Legendre decomposition can more accurately reconstruct tensors than other nonnegative tensor decomposition methods.

Bilevel learning of the Group Lasso structure

Jordan Frecon, Saverio Salzo, Massimiliano Pontil

Regression with group-sparsity penalty plays a central role in high-dimensional prediction problems. Most of existing methods require the group structure to be known a priori. In practice, this may be a too strong assumption, potentially hampering the effectiveness of the regularization method. To circumvent this issue, we present a method to estimate the group structure by means of a continuous bilevel optimization problem where the data is split into training and validation sets. Our approach relies on an approximation scheme where the lower level problem is replaced by a smooth dual forward-backward algorithm with Bregman distances. We provide guarantees regarding the convergence of the approximate procedure to the exact problem and demonstrate the well behaviour of the proposed method on synthetic experiments. Finally, a preliminary application to genes expression data is tackled with the purpose of unveiling functional groups.

SING: Symbol-to-Instrument Neural Generator

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 16k Hz. 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.

Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks

Bryan Lim

Electronic health records provide a rich source of data for machine learning methods to learn dynamic treatment responses over time. However, any direct estimation is hampered by the presence of time-dependent confounding, where actions taken are dependent on time-varying variables related to the outcome of interest. Drawing inspiration from marginal structural models, a class of methods in epidemiology which use propensity weighting to adjust for time-dependent confounders, we introduce the Recurrent Marginal Structural Network - a sequence-to-sequence architecture for forecasting a patient's expected response to a series of planned treatments. Using simulations of a state-of-the-art pharmacokinetic-pharmacodynamic (PK-PD) model of tumor growth, we demonstrate the ability of our network to accurately learn unbiased treatment responses from observational data - even under changes in the policy of treatment assignments - and performance gains over benchmarks.

Optimal Subsampling with Influence Functions

Daniel Ting, Eric Brochu

Subsampling is a common and often effective method to deal with the computational challenges of large datasets. However, for most statistical models, there is no well-motivated approach for drawing a non-uniform subsample. We show that the concept of an asymptotically linear estimator and the associated influence function leads to asymptotically optimal sampling probabilities for a wide class of popular models. This is the only tight optimality result for subsampling we are aware of as other methods only provide probabilistic error bounds or optimal rates.

Furthermore, for linear regression models, which have well-studied procedures for non-uniform subsampling, we empirically show our optimal influence function based method outperforms previous approaches even when using approximations to the optimal probabilities.

LinkNet: Relational Embedding for Scene Graph

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.

Meta-Learning MCMC Proposals

Tongzhou Wang, YI WU, Dave Moore, Stuart J. Russell

Effective implementations of sampling-based probabilistic inference often require manually constructed, model-specific proposals. Inspired by recent progresses in meta-learning for training learning agents that can generalize to unseen environments, we propose a meta-learning approach to building effective and generalizable MCMC proposals. We parametrize the proposal as a neural network to provide fast approximations to block Gibbs conditionals. The learned neural proposals g

eneralize to occurrences of common structural motifs across different models, allowing for the construction of a library of learned inference primitives that can accelerate inference on unseen models with no model-specific training required. We explore several applications including open-universe Gaussian mixture models, in which our learned proposals outperform a hand-tuned sampler, and a real-world named entity recognition task, in which our sampler yields higher final F1 scores than classical single-site Gibbs sampling.

Bayesian Adversarial Learning

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/ycxsawr}](https://tinyurl.com/ycxsawr).

Bayesian Pose Graph Optimization via Bingham Distributions and Tempered Geodesic MCMC

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.

Quadratic Decomposable Submodular Function Minimization

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.

An Improved Analysis of Alternating Minimization for Structured Multi-Response Regression

Sheng Chen, Arindam Banerjee

Multi-response linear models aggregate a set of vanilla linear models by assuming correlated noise across them, which has an unknown covariance structure. To find the coefficient vector, estimators with a joint approximation of the noise covariance are often preferred than the simple linear regression in view of their superior empirical performance, which can be generally solved by alternating-minimization type procedures. Due to the non-convex nature of such joint estimators, the theoretical justification of their efficiency is typically challenging. The existing analyses fail to fully explain the empirical observations due to the assumption of resampling on the alternating procedures, which requires access to fresh samples in each iteration. In this work, we present a resampling-free analysis for the alternating minimization algorithm applied to the multi-response regression. In particular, we focus on the high-dimensional setting of multi-response linear models with structured coefficient parameter, and the statistical error of the parameter can be expressed by the complexity measure, Gaussian width, which is related to the assumed structure. More importantly, to the best of our knowledge, our result reveals for the first time that the alternating minimization with random initialization can achieve the same performance as the well-initialized one when solving this multi-response regression problem. Experimental results support our theoretical developments.

Uniform Convergence of Gradients for Non-Convex Learning and Optimization

Dylan J. Foster, Ayush Sekhari, Karthik Sridharan

We investigate 1) the rate at which refined properties of the empirical risk---in particular, gradients---converge to their population counterparts in standard non-convex learning tasks, and 2) the consequences of this convergence for optimization. Our analysis follows the tradition of norm-based capacity control. We propose vector-valued Rademacher complexities as a simple, composable, and user-friendly tool to derive dimension-free uniform convergence bounds for gradients in non-convex learning problems. As an application of our techniques, we give a new analysis of batch gradient descent methods for non-convex generalized linear models and non-convex robust regression, showing how to use any algorithm that finds approximate stationary points to obtain optimal sample complexity, even when dimension is high or possibly infinite and multiple passes over the dataset are allowed.

Posterior Concentration for Sparse Deep Learning

Nicholas G. Polson, Veronika Ročková

We introduce Spike-and-Slab Deep Learning (SS-DL), a fully Bayesian alternative to dropout for improving generalizability of deep ReLU networks. This new type of regularization enables provable recovery of smooth input-output maps with $\{\backslash \text{sl unknown}\}$ levels of smoothness. Indeed, we show that the posterior distribution concentrates at the near minimax rate for α -Holder smooth maps, performing as well as if we knew the smoothness level α ahead of time. Our result sheds light on architecture design for deep neural networks, namely the choice of depth, width and sparsity level. These network attributes typically depend on unknown smoothness in order to be optimal. We obviate this constraint with the fully Bayes construction. As an aside, we show that SS-DL does not overfit in the sense that the posterior concentrates on smaller networks with fewer (up to the optimal number of) nodes and links. Our results provide new theoretical justifications for deep ReLU networks from a Bayesian point of view.

Sequence-to-Segment Networks for Segment Detection

Zijun Wei, Boyu Wang, Minh Hoai Nguyen, Jianming Zhang, Zhe Lin, Xiaohui Shen, Radomir Mech, Dimitris Samaras

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-auth

ors prior to requesting a name change in the electronic proceedings.

Multi-Agent Reinforcement Learning via Double Averaging Primal-Dual Optimization
Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, Mingyi Hong

Despite the success of single-agent reinforcement learning, multi-agent reinforcement learning (MARL) remains challenging due to complex interactions between agents. Motivated by decentralized applications such as sensor networks, swarm robotics, and power grids, we study policy evaluation in MARL, where agents with jointly observed state-action pairs and private local rewards collaborate to learn the value of a given policy.

In this paper, we propose a double averaging scheme, where each agent iteratively performs averaging over both space and time to incorporate neighboring gradient information and local reward information, respectively. We prove that the proposed algorithm converges to the optimal solution at a global geometric rate. In particular, such an algorithm is built upon a primal-dual reformulation of the mean squared Bellman error minimization problem, which gives rise to a decentralized convex-concave saddle-point problem. To the best of our knowledge, the proposed double averaging primal-dual optimization algorithm is the first to achieve fast finite-time convergence on decentralized convex-concave saddle-point problems.

Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Arthur Jacot, Franck Gabriel, Clement Hongler

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.

Randomized Prior Functions for Deep Reinforcement Learning

Ian Osband, John Aslanides, Albin Cassirer

Dealing with uncertainty is essential for efficient reinforcement learning. There is a growing literature on uncertainty estimation for deep learning from fixed datasets, but many of the most popular approaches are poorly-suited to sequential decision problems.

Other methods, such as bootstrap sampling, have no mechanism for uncertainty that does not come from the observed data.

We highlight why this can be a crucial shortcoming and propose a simple remedy through addition of a randomized untrainable 'prior' network to each ensemble member.

We prove that this approach is efficient with linear representations, provide simple illustrations of its efficacy with nonlinear representations and show that this approach scales to large-scale problems far better than previous attempts.

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

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, Jason D. 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

hat 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.

Uncertainty Sampling is Preconditioned Stochastic Gradient Descent on Zero-One Loss

Stephen Mussmann, Percy S. Liang

Uncertainty sampling, a popular active learning algorithm, is used to reduce the amount of data required to learn a classifier, but it has been observed in practice to converge to different parameters depending on the initialization and sometimes to even better parameters than standard training on all the data. In this work, we give a theoretical explanation of this phenomenon, showing that uncertainty sampling on a convex (e.g., logistic) loss can be interpreted as performing a preconditioned stochastic gradient step on the population zero-one loss. Experiments on synthetic and real datasets support this connection.

Negotiable Reinforcement Learning for Pareto Optimal Sequential Decision-Making
Nishant Desai, Andrew Critch, Stuart J. Russell

It is commonly believed that an agent making decisions on behalf of two or more principals who have different utility functions should adopt a Pareto optimal policy, i.e. a policy that cannot be improved upon for one principal without making sacrifices for another. Harsanyi's theorem shows that when the principals have a common prior on the outcome distributions of all policies, a Pareto optimal policy for the agent is one that maximizes a fixed, weighted linear combination of the principals' utilities. In this paper, we derive a more precise generalization for the sequential decision setting in the case of principals with different priors on the dynamics of the environment. We refer to this generalization as the Negotiable Reinforcement Learning (NRL) framework. In this more general case, the relative weight given to each principal's utility should evolve over time according to how well the agent's observations conform with that principal's prior. To gain insight into the dynamics of this new framework, we implement a simple NRL agent and empirically examine its behavior in a simple environment.

Distributed Stochastic Optimization via Adaptive SGD

Ashok Cutkosky, Róbert Busa-Fekete

Stochastic convex optimization algorithms are the most popular way to train machine learning models on large-scale data. Scaling up the training process of these models is crucial, but the most popular algorithm, Stochastic Gradient Descent (SGD), is a serial method that is surprisingly hard to parallelize. In this paper, we propose an efficient distributed stochastic optimization method by combining adaptivity with variance reduction techniques. Our analysis yields a linear speedup in the number of machines, constant memory footprint, and only a logarithmic number of communication rounds. Critically, our approach is a black-box reduction that parallelizes any serial online learning algorithm, streamlining prior analysis and allowing us to leverage the significant progress that has been made in designing adaptive algorithms. In particular, we achieve optimal convergence rates without any prior knowledge of smoothness parameters, yielding a more robust algorithm that reduces the need for hyperparameter tuning. We implement our algorithm in the Spark distributed framework and exhibit dramatic performance gains on large-scale logistic regression problems.

Smoothed Analysis of Discrete Tensor Decomposition and Assemblies of Neurons

Nima Anari, Constantinos Daskalakis, Wolfgang Maass, Christos Papadimitriou, Ami

n Saberi, Santosh Vempala

We analyze linear independence of rank one tensors produced by tensor powers of randomly perturbed vectors. This enables efficient decomposition of sums of high-order tensors. Our analysis builds upon [BCMV14] but allows for a wider range of perturbation models, including discrete ones. We give an application to recovering assemblies of neurons.

Deep State Space Models for Time Series Forecasting

Syama Sundar Rangapuram, Matthias W. Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, Tim Januschowski

We present a novel approach to probabilistic time series forecasting that combines state space models with deep learning. By parametrizing a per-time-series linear state space model with a jointly-learned recurrent neural network, our method retains desired properties of state space models such as data efficiency and interpretability, while making use of the ability to learn complex patterns from raw data offered by deep learning approaches. Our method scales gracefully from regimes where little training data is available to regimes where data from millions of time series can be leveraged to learn accurate models. We provide qualitative as well as quantitative results with the proposed method, showing that it compares favorably to the state-of-the-art.

Learning Temporal Point Processes via Reinforcement Learning

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.

GLoMo: Unsupervised Learning of Transferable Relational Graphs

Zhilin Yang, Jake Zhao, Bhuwan Dhingra, Kaiming He, William W. Cohen, Russ R. 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.

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

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.

Deep Anomaly Detection Using Geometric Transformations

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.

On Oracle-Efficient PAC RL with Rich Observations

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, Robert E. Schapire

We study the computational tractability of PAC reinforcement learning with rich observations. We present new provably sample-efficient algorithms for environments with deterministic hidden state dynamics and stochastic rich observations. These methods operate in an oracle model of computation -- accessing policy and value function classes exclusively through standard optimization primitives -- and therefore represent computationally efficient alternatives to prior algorithms that require enumeration. With stochastic hidden state dynamics, we prove that the only known sample-efficient algorithm, OLIVE, cannot be implemented in the oracle model. We also present several examples that illustrate fundamental challenges of tractable PAC reinforcement learning in such general settings.

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

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 mod

ule 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.

Towards Understanding Learning Representations: To What Extent Do Different Neural Networks Learn the Same Representation

Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, John Hopcroft

It is widely believed that learning good representations is one of the main reasons for the success of deep neural networks. Although highly intuitive, there is a lack of theory and systematic approach quantitatively characterizing what representations do deep neural networks learn. In this work, we move a tiny step towards a theory and better understanding of the representations. Specifically, we study a simpler problem: How similar are the representations learned by two networks with identical architecture but trained from different initializations. We develop a rigorous theory based on the neuron activation subspace match model. The theory gives a complete characterization of the structure of neuron activation subspace matches, where the core concepts are maximum match and simple match which describe the overall and the finest similarity between sets of neurons in two networks respectively. We also propose efficient algorithms to find the maximum match and simple matches. Finally, we conduct extensive experiments using our algorithms. Experimental results suggest that, surprisingly, representations learned by the same convolutional layers of networks trained from different initializations are not as similar as prevalently expected, at least in terms of subspace match.

Non-delusional Q-learning and value-iteration

Tyler Lu, Dale Schuurmans, Craig Boutilier

We identify a fundamental source of error in Q-learning and other forms of dynamic programming with function approximation. Delusional bias arises when the approximation architecture limits the class of expressible greedy policies. Since standard Q-updates make globally uncoordinated action choices with respect to the expressible policy class, inconsistent or even conflicting Q-value estimates can result, leading to pathological behaviour such as over/under-estimation, instability and even divergence. To solve this problem, we introduce a new notion of policy consistency and define a local backup process that ensures global consistency through the use of information sets---sets that record constraints on policies consistent with backed-up Q-values. We prove that both the model-based and model-free algorithms using this backup remove delusional bias, yielding the first known algorithms that guarantee optimal results under general conditions. These algorithms furthermore only require polynomially many information sets (from a potentially exponential support). Finally, we suggest other practical heuristics for value-iteration and Q-learning that attempt to reduce delusional bias.

An intriguing failing of convolutional neural networks and the CoordConv solution

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

m, 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.

Adversarially Robust Optimization with Gaussian Processes

Ilija Bogunovic, Jonathan Scarlett, Stefanie Jegelka, Volkan Cevher

In this paper, we consider the problem of Gaussian process (GP) optimization with an added robustness requirement: The returned point may be perturbed by an adversary, and we require the function value to remain as high as possible even after this perturbation. This problem is motivated by settings in which the underlying functions during optimization and implementation stages are different, or when one is interested in finding an entire region of good inputs rather than only a single point. We show that standard GP optimization algorithms do not exhibit the desired robustness properties, and provide a novel confidence-bound based algorithm StableOpt for this purpose. We rigorously establish the required number of samples for StableOpt to find a near-optimal point, and we complement this guarantee with an algorithm-independent lower bound. We experimentally demonstrate several potential applications of interest using real-world data sets, and we show that StableOpt consistently succeeds in finding a stable maximizer where several baseline methods fail.

Learning Hierarchical Semantic Image Manipulation through Structured Representations

Seunghoon Hong, Xinchun Yan, Thomas S. Huang, Honglak Lee

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-filling 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.

Neural Proximal Gradient Descent for Compressive Imaging

Morteza Mardani, Qingyun Sun, David Donoho, Vardan Papayan, Hatef Monajemi, Shreyas Vasanthawala, John Pauly

Recovering high-resolution images from limited sensory data typically leads to a serious ill-posed inverse problem, demanding inversion algorithms that effectively

ely 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.

Power-law efficient neural codes provide general link between perceptual bias and discriminability

Michael Morais, Jonathan W. Pillow

Recent work in theoretical neuroscience has shown that information-theoretic "efficient" neural codes, which allocate neural resources to maximize the mutual information between stimuli and neural responses, give rise to a lawful relationship between perceptual bias and discriminability that is observed across a wide variety of psychophysical tasks in human observers (Wei & Stocker 2017). Here we generalize these results to show that the same law arises under a much larger family of optimal neural codes, introducing a unifying framework that we call power-law efficient coding. Specifically, we show that the same lawful relationship between bias and discriminability arises whenever Fisher information is allocated proportional to any power of the prior distribution. This family includes neural codes that are optimal for minimizing L_p error for any p , indicating that the lawful relationship observed in human psychophysical data does not require information-theoretically optimal neural codes. Furthermore, we derive the exact constant of proportionality governing the relationship between bias and discriminability for different power laws (which includes information-theoretically optimal codes, where the power is 2, and so-called discrimax codes, where power is $1/2$), and different choices of optimal decoder. As a bonus, our framework provides new insights into "anti-Bayesian" perceptual biases, in which percepts are biased away from the center of mass of the prior. We derive an explicit formula that clarifies precisely which combinations of neural encoder and decoder can give rise to such biases.

Stochastic Nonparametric Event-Tensor Decomposition

Shandian Zhe, Yishuai Du

Tensor decompositions are fundamental tools for multiway data analysis. Existing approaches, however, ignore the valuable temporal information along with data, or simply discretize them into time steps so that important temporal patterns are easily missed. Moreover, most methods are limited to multilinear decomposition forms, and hence are unable to capture intricate, nonlinear relationships in data. To address these issues, we formulate event-tensors, to preserve the complete temporal information for multiway data, and propose a novel Bayesian nonparametric decomposition model. Our model can (1) fully exploit the time stamps to capture the critical, causal/triggering effects between the interaction events, (2) flexibly estimate the complex relationships between the entities in tensor modes, and (3) uncover hidden structures from their temporal interactions. For scalable inference, we develop a doubly stochastic variational Expectation-Maximization algorithm to conduct an online decomposition. Evaluations on both synthetic and real-world datasets show that our model not only improves upon the predictive performance of existing methods, but also discovers interesting clusters underlying the data.

A Smoother Way to Train Structured Prediction Models

Venkata Krishna Pillutla, Vincent Roulet, Sham M. Kakade, Zaid Harchaoui

We present a framework to train a structured prediction model by performing smoothing on the inference algorithm it builds upon. Smoothing overcomes the non-smoothness inherent to the maximum margin structured prediction objective, and paves the way for the use of fast primal gradient-based optimization algorithms. We illustrate the proposed framework by developing a novel primal incremental optimization algorithm for the structural support vector machine. The proposed algorithm blends an extrapolation scheme for acceleration and an adaptive smoothing scheme and builds upon the stochastic variance-reduced gradient algorithm. We establish its worst-case global complexity bound and study several practical variants. We present experimental results on two real-world problems, namely named entity recognition and visual object localization. The experimental results show that the proposed framework allows us to build upon efficient inference algorithms to develop large-scale optimization algorithms for structured prediction which can achieve competitive performance on the two real-world problems.

Evolutionary Stochastic Gradient Descent for Optimization of Deep Neural Networks

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.

On Coresets for Logistic Regression

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, David Woodruff

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Assessing the Scalability of Biologically-Motivated Deep Learning Algorithms and Architectures

Sergey Bartunov, Adam Santoro, Blake Richards, Luke Marris, Geoffrey E. Hinton, Timothy Lillicrap

The backpropagation of error algorithm (BP) is impossible to implement in a real brain. The recent success of deep networks in machine learning and AI, however, has inspired proposals for understanding how the brain might learn across multiple layers, and hence how it might approximate BP. As of yet, none of these proposals have been rigorously evaluated on tasks where BP-guided deep learning has proved critical, or in architectures more structured than simple fully-connected networks. Here we present results on scaling up biologically motivated models of deep learning on datasets which need deep networks with appropriate architectures to achieve good performance. We present results on the MNIST, CIFAR-10, and ImageNet datasets and explore variants of target-propagation (TP) and feedback alignment (FA) algorithms, and explore performance in both fully- and locally-connected architectures. We also introduce weight-transport-free variants of difference target propagation (DTP) modified to remove backpropagation from the penultimate layer. Many of these algorithms perform well for MNIST, but for CIFAR and

ImageNet we find that TP and FA variants perform significantly worse than BP, especially for networks composed of locally connected units, opening questions about whether new architectures and algorithms are required to scale these approaches. Our results and implementation details help establish baselines for biologically motivated deep learning schemes going forward.

3D-Aware Scene Manipulation via Inverse Graphics

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.

Learn What Not to Learn: Action Elimination with Deep Reinforcement Learning

Tom Zahavy, Matan Haroush, Nadav Merlis, Daniel J. Mankowitz, Shie Mannor

Learning how to act when there are many available actions in each state is a challenging task for Reinforcement Learning (RL) agents, especially when many of the actions are redundant or irrelevant. In such cases, it is easier to learn which actions not to take. In this work, we propose the Action-Elimination Deep Q-Network (AE-DQN) architecture that combines a Deep RL algorithm with an Action Elimination Network (AEN) that eliminates sub-optimal actions. The AEN is trained to predict invalid actions, supervised by an external elimination signal provided by the environment. Simulations demonstrate a considerable speedup and robustness over vanilla DQN in text-based games with over a thousand discrete actions.

Connecting Optimization and Regularization Paths

Arun Suggala, Adarsh Prasad, Pradeep K. Ravikumar

We study the implicit regularization properties of optimization techniques by explicitly connecting their optimization paths to the regularization paths of 'corresponding' regularized problems. This surprising connection shows that iterates of optimization techniques such as gradient descent and mirror descent are 'close' to solutions of appropriately regularized objectives. While such a tight connection between optimization and regularization is of independent intellectual interest, it also has important implications for machine learning: we can port results from regularized estimators to optimization, and vice versa. We investigate one key consequence, that borrows from the well-studied analysis of regularized estimators, to then obtain tight excess risk bounds of the iterates generated by optimization techniques.

A Theory-Based Evaluation of Nearest Neighbor Models Put Into Practice

Hendrik Fichtenberger, Dennis Rohde

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

MetaReg: Towards Domain Generalization using Meta-Regularization

Yogesh Balaji, Swami Sankaranarayanan, Rama Chellappa

Training models that generalize to new domains at test time is a problem of fundamental importance in machine learning. In this work, we encode this notion of domain generalization using a novel regularization function. We pose the problem of finding such a regularization function in a Learning to Learn (or) meta-learning framework. The objective of domain generalization is explicitly modeled by learning a regularizer that makes the model trained on one domain to perform well on another domain. Experimental validations on computer vision and natural language datasets indicate that our method can learn regularizers that achieve good cross-domain generalization.

Mirrored Langevin Dynamics

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, Volkan Cevher

We consider the problem of sampling from constrained distributions, which has posed significant challenges to both non-asymptotic analysis and algorithmic design. We propose a unified framework, which is inspired by the classical mirror descent, to derive novel first-order sampling schemes. We prove that, for a general target distribution with strongly convex potential, our framework implies the existence of a first-order algorithm achieving $O(\epsilon^{-2}d)$ convergence, suggesting that the state-of-the-art $O(\epsilon^{-6}d^5)$ can be vastly improved. With the important Latent Dirichlet Allocation (LDA) application in mind, we specialize our algorithm to sample from Dirichlet posteriors, and derive the first non-asymptotic $O(\epsilon^{-2}d^2)$ rate for first-order sampling. We further extend our framework to the mini-batch setting and prove convergence rates when only stochastic gradients are available. Finally, we report promising experimental results for LDA on real datasets.

Multivariate Convolutional Sparse Coding for Electromagnetic Brain Signals

Tom Dupré la Tour, Thomas Moreau, Mainak Jas, Alexandre Gramfort

Frequency-specific patterns of neural activity are traditionally interpreted as sustained rhythmic oscillations, and related to cognitive mechanisms such as attention, high level visual processing or motor control. While alpha waves (8--12 Hz) are known to closely resemble short sinusoids, and thus are revealed by Fourier analysis or wavelet transforms, there is an evolving debate that electromagnetic neural signals are composed of more complex waveforms that cannot be analyzed by linear filters and traditional signal representations. In this paper, we propose to learn dedicated representations of such recordings using a multivariate convolutional sparse coding (CSC) algorithm. Applied to electroencephalography (EEG) or magnetoencephalography (MEG) data, this method is able to learn not only prototypical temporal waveforms, but also associated spatial patterns so their origin can be localized in the brain. Our algorithm is based on alternated minimization and a greedy coordinate descent solver that leads to state-of-the-art running time on long time series. To demonstrate the implications of this method, we apply it to MEG data and show that it is able to recover biological artifacts. More remarkably, our approach also reveals the presence of non-sinusoidal mu-shaped patterns, along with their topographic maps related to the somatosensory cortex.

Complex Gated Recurrent Neural Networks

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.

Active Matting

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.

Scalable End-to-End Autonomous Vehicle Testing via Rare-event Simulation
Matthew O'Kelly, Aman Sinha, Hongseok Namkoong, Russ Tedrake, John C. 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.

Improving Explorability in Variational Inference with Annealed Variational Objectives
Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, Aaron C. Courville
Despite the advances in the representational capacity of approximate distributions for variational inference, the optimization process can still limit the density that is ultimately learned. We demonstrate the drawbacks of biasing the true posterior to be unimodal, and introduce Annealed Variational Objectives (AVO) into the training of hierarchical variational methods. Inspired by Annealed Importance Sampling, the proposed method facilitates learning by incorporating energy tempering into the optimization objective. In our experiments, we demonstrate our method's robustness to deterministic warm up, and the benefits of encouraging exploration in the latent space.

Learning Loop Invariants for Program Verification
Xujie Si, Hanjun Dai, Mukund Raghothaman, Mayur Naik, Le Song
A fundamental problem in program verification concerns inferring loop invariants. The problem is undecidable and even practical instances are challenging. Inspired by how human experts construct loop invariants, we propose a reasoning framework Code2Inv that constructs the solution by multi-step decision making and queuing an external program graph memory block. By training with reinforcement learning, Code2Inv captures rich program features and avoids the need for ground truth solutions as supervision. Compared to previous learning tasks in domains with graph-structured data, it addresses unique challenges, such as a binary objective function and an extremely sparse reward that is given by an automated theorem prover only after the complete loop invariant is proposed. We evaluate Code2Inv

v on a suite of 133 benchmark problems and compare it to three state-of-the-art systems. It solves 106 problems compared to 73 by a stochastic search-based system, 77 by a heuristic search-based system, and 100 by a decision tree learning-based system. Moreover, the strategy learned can be generalized to new programs: compared to solving new instances from scratch, the pre-trained agent is more sample efficient in finding solutions.

Faster Online Learning of Optimal Threshold for Consistent F-measure Optimization

Xiaoxuan Zhang, Mingrui Liu, Xun Zhou, Tianbao Yang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning Others' Intentional Models in Multi-Agent Settings Using Interactive POMDPs

Yanlin Han, Piotr Gmytrasiewicz

Interactive partially observable Markov decision processes (I-POMDPs) provide a principled framework for planning and acting in a partially observable, stochastic and multi-agent environment. It extends POMDPs to multi-agent settings by including models of other agents in the state space and forming a hierarchical belief structure. In order to predict other agents' actions using I-POMDPs, we propose an approach that effectively uses Bayesian inference and sequential Monte Carlo sampling to learn others' intentional models which ascribe to them beliefs, preferences and rationality in action selection. Empirical results show that our algorithm accurately learns models of the other agent and has superior performance than methods that use subintentional models. Our approach serves as a generalized Bayesian learning algorithm that learns other agents' beliefs, strategy levels, and transition, observation and reward functions. It also effectively mitigates the belief space complexity due to the nested belief hierarchy.

Online Structure Learning for Feed-Forward and Recurrent Sum-Product Networks

Agastya Kalra, Abdullah Rashwan, Wei-Shou Hsu, Pascal Poupart, Prashant Doshi, Georgios Trimonias

Sum-product networks have recently emerged as an attractive representation due to their dual view as a special type of deep neural network with clear semantics and a special type of probabilistic graphical model for which inference is always tractable. Those properties follow from some conditions (i.e., completeness and decomposability) that must be respected by the structure of the network. As a result, it is not easy to specify a valid sum-product network by hand and therefore structure learning techniques are typically used in practice. This paper describes a new online structure learning technique for feed-forward and recurrent SPNs. The algorithm is demonstrated on real-world datasets with continuous features for which it is not clear what network architecture might be best, including sequence datasets of varying length.

Balanced Policy Evaluation and Learning

Nathan Kallus

We present a new approach to the problems of evaluating and learning personalized decision policies from observational data of past contexts, decisions, and outcomes. Only the outcome of the enacted decision is available and the historical policy is unknown. These problems arise in personalized medicine using electronic health records and in internet advertising. Existing approaches use inverse propensity weighting (or, doubly robust versions) to make historical outcome (or, residual) data look like it were generated by a new policy being evaluated or learned. But this relies on a plug-in approach that rejects data points with a decision that disagrees with the new policy, leading to high variance estimates and ineffective learning. We propose a new, balance-based approach that too makes the data look like the new policy but does so directly by finding weights that op

timize for balance between the weighted data and the target policy in the given, finite sample, which is equivalent to minimizing worst-case or posterior conditional mean square error. Our policy learner proceeds as a two-level optimization problem over policies and weights. We demonstrate that this approach markedly outperforms existing ones both in evaluation and learning, which is unsurprising given the wider support of balance-based weights. We establish extensive theoretical consistency guarantees and regret bounds that support this empirical success.

Visual Memory for Robust Path Following

Ashish Kumar, Saurabh Gupta, David Fouhey, Sergey Levine, Jitendra Malik

Humans routinely retrace a path in a novel environment both forwards and backwards despite uncertainty in their motion. In this paper, we present an approach for doing so. Given a demonstration of a path, a first network generates an abstraction of the path. Equipped with this abstraction, a second network then observes the world and decides how to act in order to retrace the path under noisy actuation and a changing environment. The two networks are optimized end-to-end at training time. We evaluate the method in two realistic simulators, performing path following both forwards and backwards. Our experiments show that our approach outperforms both a classical approach to solving this task as well as a number of other baselines.

Representation Learning of Compositional Data

Marta Avalos, Richard Nock, Cheng Soon Ong, Julien Rouar, Ke Sun

We consider the problem of learning a low dimensional representation for compositional data. Compositional data consists of a collection of nonnegative data that sum to a constant value. Since the parts of the collection are statistically dependent, many standard tools cannot be directly applied. Instead, compositional data must be first transformed before analysis. Focusing on principal component analysis (PCA), we propose an approach that allows low dimensional representation learning directly from the original data. Our approach combines the benefits of the log-ratio transformation from compositional data analysis and exponential family PCA. A key tool in its derivation is a generalization of the scaled Bregman theorem, that relates the perspective transform of a Bregman divergence to the Bregman divergence of a perspective transform and a remainder conformal divergence. Our proposed approach includes a convenient surrogate (upper bound) loss of the exponential family PCA which has an easy to optimize form. We also derive the corresponding form for nonlinear autoencoders. Experiments on simulated data and microbiome data show the promise of our method.

How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective

Lei Wu, Chao Ma, Weinan E

The question of which global minima are accessible by a stochastic gradient descent (SGD) algorithm with specific learning rate and batch size is studied from the perspective of dynamical stability. The concept of non-uniformity is introduced, which, together with sharpness, characterizes the stability property of a global minimum and hence the accessibility of a particular SGD algorithm to that global minimum. In particular, this analysis shows that learning rate and batch size play different roles in minima selection. Extensive empirical results seem to correlate well with the theoretical findings and provide further support to these claims.

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

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.

A Bayes-Sard Cubature Method

Toni Karvonen, Chris J. Oates, Simo Sarkka

This paper focusses on the formulation of numerical integration as an inferential task. To date, research effort has largely focussed on the development of Bayesian cubature, whose distributional output provides uncertainty quantification for the integral. However, the point estimators associated to Bayesian cubature can be inaccurate and acutely sensitive to the prior when the domain is high-dimensional. To address these drawbacks we introduce Bayes-Sard cubature, a probabilistic framework that combines the flexibility of Bayesian cubature with the robustness of classical cubatures which are well-established. This is achieved by considering a Gaussian process model for the integrand whose mean is a parametric regression model, with an improper prior on each regression coefficient. The features in the regression model consist of test functions which are guaranteed to be exactly integrated, with remaining degrees of freedom afforded to the non-parametric part. The asymptotic convergence of the Bayes-Sard cubature method is established and the theoretical results are numerically verified. In particular, we report two orders of magnitude reduction in error compared to Bayesian cubature in the context of a high-dimensional financial integral.

Learning to Infer Graphics Programs from Hand-Drawn Images

Kevin Ellis, Daniel Ritchie, Armando Solar-Lezama, Josh Tenenbaum

We introduce a model that learns to convert simple hand drawings into graphics programs written in a subset of \backslash LaTeX. The model combines techniques from deep learning and program synthesis. We learn a convolutional neural network that proposes plausible drawing primitives that explain an image. These drawing primitives are a specification (spec) of what the graphics program needs to draw. We learn a model that uses program synthesis techniques to recover a graphics program from that spec. These programs have constructs like variable bindings, iterative loops, or simple kinds of conditionals. With a graphics program in hand, we can correct errors made by the deep network and extrapolate drawings.

Neural Guided Constraint Logic Programming for Program Synthesis

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.

Overcoming Language Priors in Visual Question Answering with Adversarial Regularization

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.

New Insight into Hybrid Stochastic Gradient Descent: Beyond With-Replacement Sampling and Convexity

Pan Zhou, Xiaotong Yuan, Jiashi Feng

As an incremental-gradient algorithm, the hybrid stochastic gradient descent (HSGD) enjoys merits of both stochastic and full gradient methods for finite-sum minimization problem. However, the existing rate-of-convergence analysis for HSGD is made under with-replacement sampling (WRS) and is restricted to convex problems. It is not clear whether HSGD still carries these advantages under the common practice of without-replacement sampling (WoRS) for non-convex problems. In this paper, we affirmatively answer this open question by showing that under WoRS and for both convex and non-convex problems, it is still possible for HSGD (with constant step-size) to match full gradient descent in rate of convergence, while maintaining comparable sample-size-independent incremental first-order oracle complexity to stochastic gradient descent. For a special class of finite-sum problems with linear prediction models, our convergence results can be further improved in some cases. Extensive numerical results confirm our theoretical affirmation and demonstrate the favorable efficiency of WoRS-based HSGD.

Variational PDEs for Acceleration on Manifolds and Application to Diffeomorphisms

Ganesh Sundaramoorthi, Anthony Yezzi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis

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.

Learning to Solve SMT Formulas

Mislav Balunovic, Pavol Bielik, Martin Vechev

We present a new approach for learning to solve SMT formulas. We phrase the chal

length 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.

Learning to Repair Software Vulnerabilities with Generative Adversarial Networks
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.

Algorithmic Linearly Constrained Gaussian Processes
Markus Lange-Hegemann

We algorithmically construct multi-output Gaussian process priors which satisfy linear differential equations. Our approach attempts to parametrize all solutions of the equations using Gröbner bases. If successful, a push forward Gaussian process along the parameterization is the desired prior. We consider several examples from physics, geomathematics and control, among them the full inhomogeneous system of Maxwell's equations. By bringing together stochastic learning and computational algebra in a novel way, we combine noisy observations with precise algebraic computations.

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

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.

Universal Growth in Production Economies
Simina Branzei, Ruta Mehta, Noam Nisan

We study a simple variant of the von Neumann model of an expanding economy, in which multiple producers make goods according to their production function. The players trade their goods at the market and then use the bundles received as input

ts for the production in the next round. The decision that players have to make is how to invest their money (i.e. bids) in each round.

Neural Ordinary Differential Equations

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, David K. 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.

MetaAnchor: Learning to Detect Objects with Customized Anchors

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.

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

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.

On Controllable Sparse Alternatives to Softmax

Anirban Laha, Saneem Ahmed Chemmengath, Priyanka Agrawal, Mitesh Khapra, Karthik Sankaranarayanan, Harish G. Ramaswamy

Converting an n-dimensional vector to a probability distribution over n objects is a commonly used component in many machine learning tasks like multiclass classification, multilabel classification, attention mechanisms etc. For this, several probability mapping functions have been proposed and employed in literature such as softmax, sum-normalization, spherical softmax, and sparsemax, but there is very little understanding in terms how they relate with each other. Further, none of the above formulations offer an explicit control over the degree of spars

ity. To address this, we develop a unified framework that encompasses all these formulations as special cases. This framework ensures simple closed-form solutions and existence of sub-gradients suitable for learning via backpropagation. Within this framework, we propose two novel sparse formulations, *sparsegen-lin* and *sparsehourglass*, that seek to provide a control over the degree of desired sparsity. We further develop novel convex loss functions that help induce the behavior of aforementioned formulations in the multilabel classification setting, showing improved performance. We also demonstrate empirically that the proposed formulations, when used to compute attention weights, achieve better or comparable performance on standard seq2seq tasks like neural machine translation and abstractive summarization.

Gaussian Process Conditional Density Estimation

Vincent Dutoit, Hugh Salimbeni, James Hensman, Marc Deisenroth

Conditional Density Estimation (CDE) models deal with estimating conditional distributions. The conditions imposed on the distribution are the inputs of the model. CDE is a challenging task as there is a fundamental trade-off between model complexity, representational capacity and overfitting. In this work, we propose to extend the model's input with latent variables and use Gaussian processes (GP) to map this augmented input onto samples from the conditional distribution. Our Bayesian approach allows for the modeling of small datasets, but we also provide the machinery for it to be applied to big data using stochastic variational inference. Our approach can be used to model densities even in sparse data regions, and allows for sharing learned structure between conditions. We illustrate the effectiveness and wide-reaching applicability of our model on a variety of real-world problems, such as spatio-temporal density estimation of taxi drop-offs, non-Gaussian noise modeling, and few-shot learning on omniglot images.

Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images

Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, Cristian Sminicescu

We present MubyNet -- a feed-forward, multitask, bottom up system for the integrated localization, as well as 3d pose and shape estimation, of multiple people in monocular images. The challenge is the formal modeling of the problem that intrinsically requires discrete and continuous computation, e.g. grouping people vs. predicting 3d pose. The model identifies human body structures (joints and limbs) in images, groups them based on 2d and 3d information fused using learned scoring functions, and optimally aggregates such responses into partial or complete 3d human skeleton hypotheses under kinematic tree constraints, but without knowing in advance the number of people in the scene and their visibility relations. We design a multi-task deep neural network with differentiable stages where the person grouping problem is formulated as an integer program based on learned body part scores parameterized by both 2d and 3d information. This avoids suboptimality resulting from separate 2d and 3d reasoning, with grouping performed based on the combined representation. The final stage of 3d pose and shape prediction is based on a learned attention process where information from different human body parts is optimally integrated. State-of-the-art results are obtained in large scale datasets like Human3.6M and Panoptic, and qualitatively by reconstructing the 3d shape and pose of multiple people, under occlusion, in difficult monocular images.

Learning Attentional Communication for Multi-Agent Cooperation

Jiechuan Jiang, Zongqing Lu

Communication could potentially be an effective way for multi-agent cooperation.

However, information sharing among all agents or in predefined communication architectures that existing methods adopt can be problematic. When there is a large number of agents, agents cannot differentiate valuable information that helps cooperative decision making from globally shared information. Therefore, communication barely helps, and could even impair the learning of multi-agent cooperation. Predefined communication architectures, on the other hand, restrict communic

ation among agents and thus restrain potential cooperation. To tackle these difficulties, in this paper, we propose an attentional communication model that learns when communication is needed and how to integrate shared information for cooperative decision making. Our model leads to efficient and effective communication for large-scale multi-agent cooperation. Empirically, we show the strength of our model in a variety of cooperative scenarios, where agents are able to develop more coordinated and sophisticated strategies than existing methods.

Speaker-Follower Models for Vision-and-Language Navigation

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.

Training DNNs with Hybrid Block Floating Point

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 the m. 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.

Coupled Variational Bayes via Optimization Embedding

Bo Dai, Hanjun Dai, Niao He, Weiyang Liu, Zhen Liu, Jianshu Chen, Lin Xiao, Le Song

Variational inference plays a vital role in learning graphical models, especially on large-scale datasets. Much of its success depends on a proper choice of auxiliary distribution class for posterior approximation. However, how to pursue an auxiliary distribution class that achieves both good approximation ability and computation efficiency remains a core challenge. In this paper, we proposed coupled variational Bayes which exploits the primal-dual view of the ELBO with the variational distribution class generated by an optimization procedure, which is termed optimization embedding. This flexible function class couples the variatio

nal distribution with the original parameters in the graphical models, allowing end-to-end learning of the graphical models by back-propagation through the variational distribution. Theoretically, we establish an interesting connection to gradient flow and demonstrate the extreme flexibility of this implicit distribution family in the limit sense. Empirically, we demonstrate the effectiveness of the proposed method on multiple graphical models with either continuous or discrete latent variables comparing to state-of-the-art methods.

Multi-domain Causal Structure Learning in Linear Systems

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, Kun Zhang

We study the problem of causal structure learning in linear systems from observational data given in multiple domains, across which the causal coefficients and/or the distribution of the exogenous noises may vary. The main tool used in our approach is the principle that in a causally sufficient system, the causal modules, as well as their included parameters, change independently across domains. We first introduce our approach for finding causal direction in a system comprising two variables and propose efficient methods for identifying causal direction.

Then we generalize our methods to causal structure learning in networks of variables. Most of previous work in structure learning from multi-domain data assume that certain types of invariance are held in causal modules across domains. Our approach unifies the idea in those works and generalizes to the case that there is no such invariance across the domains. Our proposed methods are generally capable of identifying causal direction from fewer than ten domains. When the invariance property holds, two domains are generally sufficient.

Policy Optimization via Importance Sampling

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, Marcello Restelli

Policy optimization is an effective reinforcement learning approach to solve continuous control tasks. Recent achievements have shown that alternating online and offline optimization is a successful choice for efficient trajectory reuse. However, deciding when to stop optimizing and collect new trajectories is non-trivial, as it requires to account for the variance of the objective function estimate. In this paper, we propose a novel, model-free, policy search algorithm, POIS, applicable in both action-based and parameter-based settings. We first derive a high-confidence bound for importance sampling estimation; then we define a surrogate objective function, which is optimized offline whenever a new batch of trajectories is collected. Finally, the algorithm is tested on a selection of continuous control tasks, with both linear and deep policies, and compared with state-of-the-art policy optimization methods.

Task-Driven Convolutional Recurrent Models of the Visual System

Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J. DiCarlo, Daniel L. 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.

Contrastive Learning from Pairwise Measurements

Yi Chen, Zhuoran Yang, Yuchen Xie, Zhaoran Wang

Learning from pairwise measurements naturally arises from many applications, such as rank aggregation, ordinal embedding, and crowdsourcing. However, most existing models and algorithms are susceptible to potential model misspecification. In this paper, we study a semiparametric model where the pairwise measurements follow a natural exponential family distribution with an unknown base measure. Such a semiparametric model includes various popular parametric models, such as the Bradley-Terry-Luce model and the paired cardinal model, as special cases. To estimate this semiparametric model without specifying the base measure, we propose a data augmentation technique to create virtual examples, which enables us to define a contrastive estimator. In particular, we prove that such a contrastive estimator is invariant to model misspecification within the natural exponential family, and moreover, attains the optimal statistical rate of convergence up to a logarithmic factor. We provide numerical experiments to corroborate our theory.

Regret Bounds for Online Portfolio Selection with a Cardinality Constraint

Shinji Ito, Daisuke Hatano, Hanna Sumita, Akihiro Yabe, Takuro Fukunaga, Naonori Kakimura, Ken-Ichi Kawarabayashi

Online portfolio selection is a sequential decision-making problem in which a learner repetitively selects a portfolio over a set of assets, aiming to maximize long-term return. In this paper, we study the problem with the cardinality constraint that the number of assets in a portfolio is restricted to be at most k , and consider two scenarios: (i) in the full-feedback setting, the learner can observe price relatives (rates of return to cost) for all assets, and (ii) in the bandit-feedback setting, the learner can observe price relatives only for invested assets. We propose efficient algorithms for these scenarios that achieve sublinear regrets. We also provide regret (statistical) lower bounds for both scenarios which nearly match the upper bounds when k is a constant. In addition, we give a computational lower bound which implies that no algorithm maintains both computational efficiency, as well as a small regret upper bound.

Hunting for Discriminatory Proxies in Linear Regression Models

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.

Entropy and mutual information in models of deep neural networks

Marylou Gabri  , Andre Manoel, Cl  ment Luneau, Jean Barbier, Nicolas Macris, Florent Krzakala, Lenka Zdeborov  

We examine a class of stochastic deep learning models with a tractable method to compute information-theoretic quantities. Our contributions are three-fold: (i) We show how entropies and mutual informations can be derived from heuristic statistical physics methods, under the assumption that weight matrices are independent and orthogonally-invariant. (ii) We extend particular cases in which this result is known to be rigorously exact by providing a proof for two-layers networks with Gaussian random weights, using the recently introduced adaptive interpolation method. (iii) We propose an experiment framework with generative models of

synthetic datasets, on which we train deep neural networks with a weight constraint designed so that the assumption in (i) is verified during learning. We study the behavior of entropies and mutual information throughout learning and conclude that, in the proposed setting, the relationship between compression and generalization remains elusive.

Paraphrasing Complex Network: Network Compression via Factor Transfer

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.

A Simple Cache Model for Image Recognition

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.

Learning Attractor Dynamics for Generative Memory

Yan Wu, Gregory Wayne, Karol Gregor, Timothy Lillicrap

A central challenge faced by memory systems is the robust retrieval of a stored pattern in the presence of interference due to other stored patterns and noise. A theoretically well-founded solution to robust retrieval is given by attractor dynamics, which iteratively cleans up patterns during recall. However, incorporating attractor dynamics into modern deep learning systems poses difficulties: attractor basins are characterised by vanishing gradients, which are known to make training neural networks difficult. In this work, we exploit recent advances in variational inference and avoid the vanishing gradient problem by training a generative distributed memory with a variational lower-bound-based Lyapunov function. The model is minimalistic with surprisingly few parameters. Experiments show it converges to correct patterns upon iterative retrieval and achieves competitive performance as both a memory model and a generative model.

Quadrature-based features for kernel approximation

Marina Munkhoeva, Yermek Kapushev, Evgeny Burnaev, Ivan Oseledets

We consider the problem of improving kernel approximation via randomized feature maps. These maps arise as Monte Carlo approximation to integral representations of kernel functions and scale up kernel methods for larger datasets. Based on an efficient numerical integration technique, we propose a unifying approach that reinterprets the previous random features methods and extends to better estimates of the kernel approximation. We derive the convergence behavior and conduct an extensive empirical study that supports our hypothesis.

Efficient Algorithms for Non-convex Isotonic Regression through Submodular Optimization

Francis Bach

We consider the minimization of submodular functions subject to ordering constraints. We show that this potentially non-convex optimization problem can be cast as a convex optimization problem on a space of uni-dimensional measures, with ordering constraints corresponding to first-order stochastic dominance. We propose new discretization schemes that lead to simple and efficient algorithms based on zero-th, first, or higher order oracles; these algorithms also lead to improvements without isotonic constraints. Finally, our experiments show that non-convex loss functions can be much more robust to outliers for isotonic regression, while still being solvable in polynomial time.

Deep Neural Nets with Interpolating Function as Output Activation

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>.

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

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.

Image Inpainting via Generative Multi-column Convolutional Neural Networks

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.

Clustering Redemption-Beyond the Impossibility of Kleinberg's Axioms

Vincent Cohen-Addad, Varun Kanade, Frederik Mallmann-Trenn

Kleinberg (2002) stated three axioms that any clustering procedure should satisfy and showed there is no clustering procedure that simultaneously satisfies all three. One of these, called the consistency axiom, requires that when the data is modified in a helpful way, i.e. if points in the same cluster are made more similar and those in different ones made less similar, the algorithm should output the same clustering. To circumvent this impossibility result, research has focused on considering clustering procedures that have a clustering quality measure (or a cost) and showing that a modification of Kleinberg's axioms that takes cost into account lead to feasible clustering procedures. In this work, we take a different approach, based on the observation that the consistency axiom fails to be satisfied when the "correct" number of clusters changes. We modify this axiom by making use of cost functions to determine the correct number of clusters, and require that consistency holds only if the number of clusters remains unchanged. We show that single linkage satisfies the modified axioms, and if the input is well-clusterable, some popular procedures such as k-means also satisfy the axioms, taking a step towards explaining the success of these objective functions for guiding the design of algorithms.

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

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.

Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes

Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, Yaniv Plan

We prove that $\Omega(k d^2 / \epsilon^2)$ samples are necessary and sufficient for learning a mixture of k Gaussians in \mathbb{R}^d , up to error ϵ in total variation distance. This improves both the known upper bounds and lower bounds for this problem. For mixt

ures of axis-aligned Gaussians, we show that $O(kd / \epsilon^2)$ samples suffice, matching a known lower bound.

Object-Oriented Dynamics Predictor

Guangxiang Zhu, Zhiao Huang, Chongjie Zhang

Generalization has been one of the major challenges for learning dynamics models in model-based reinforcement learning. However, previous work on action-conditioned dynamics prediction focuses on learning the pixel-level motion and thus does not generalize well to novel environments with different object layouts. In this paper, we present a novel object-oriented framework, called object-oriented dynamics predictor (OODP), which decomposes the environment into objects and predicts the dynamics of objects conditioned on both actions and object-to-object relations. It is an end-to-end neural network and can be trained in an unsupervised manner. To enable the generalization ability of dynamics learning, we design a novel CNN-based relation mechanism that is class-specific (rather than object-specific) and exploits the locality principle. Empirical results show that OODP significantly outperforms previous methods in terms of generalization over novel environments with various object layouts. OODP is able to learn from very few environments and accurately predict dynamics in a large number of unseen environments. In addition, OODP learns semantically and visually interpretable dynamics models.

Fast Rates of ERM and Stochastic Approximation: Adaptive to Error Bound Conditions

Mingrui Liu, Xiaoxuan Zhang, Lijun Zhang, Rong Jin, Tianbao Yang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Adversarial Multiple Source Domain Adaptation

Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P. Costeira, Geoffrey J. 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.

To Trust Or Not To Trust A Classifier

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 $\{\text{it trust score}\}$, 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 distribution

al 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.

Deep Reinforcement Learning of Marked Temporal Point Processes

Utkarsh Upadhyay, Abir De, Manuel Gomez Rodriguez

In a wide variety of applications, humans interact with a complex environment by means of asynchronous stochastic discrete events in continuous time. Can we design online interventions that will help humans achieve certain goals in such asynchronous setting? In this paper, we address the above problem from the perspective of deep reinforcement learning of marked temporal point processes, where both the actions taken by an agent and the feedback it receives from the environment are asynchronous stochastic discrete events characterized using marked temporal point processes. In doing so, we define the agent's policy using the intensity and mark distribution of the corresponding process and then derive a flexible policy gradient method, which embeds the agent's actions and the feedback it receives into real-valued vectors using deep recurrent neural networks. Our method does not make any assumptions on the functional form of the intensity and mark distribution of the feedback and it allows for arbitrarily complex reward functions. We apply our methodology to two different applications in viral marketing and personalized teaching and, using data gathered from Twitter and Duolingo, we show that it may be able to find interventions to help marketers and learners achieve their goals more effectively than alternatives.

Learning to Play With Intrinsically-Motivated, Self-Aware Agents

Nick Haber, Damian Mrowca, Stephanie Wang, Li F. Fei-Fei, Daniel L. Yamins

Infants are experts at playing, with an amazing ability to generate novel structured behaviors in unstructured environments that lack clear extrinsic reward signals. We seek to mathematically formalize these abilities using a neural network that implements curiosity-driven intrinsic motivation. Using a simple but ecologically naturalistic simulated environment in which an agent can move and interact with objects it sees, we propose a "world-model" network that learns to predict the dynamic consequences of the agent's actions. Simultaneously, we train a separate explicit "self-model" that allows the agent to track the error map of its world-model. It then uses the self-model to adversarially challenge the developing world-model. We demonstrate that this policy causes the agent to explore novel and informative interactions with its environment, leading to the generation of a spectrum of complex behaviors, including ego-motion prediction, object attention, and object gathering. Moreover, the world-model that the agent learns supports improved performance on object dynamics prediction, detection, localization and recognition tasks. Taken together, our results are initial steps toward creating flexible autonomous agents that self-supervise in realistic physical environments.

Distributed Learning without Distress: Privacy-Preserving Empirical Risk Minimization

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.

Continuous-time Value Function Approximation in Reproducing Kernel Hilbert Spaces

Motoya Ohnishi, Masahiro Yukawa, Mikael Johansson, Masashi Sugiyama

Motivated by the success of reinforcement learning (RL) for discrete-time tasks such as AlphaGo and Atari games, there has been a recent surge of interest in using RL for continuous-time control of physical systems (cf. many challenging tasks in OpenAI Gym and DeepMind Control Suite).

Since discretization of time is susceptible to error, it is methodologically more desirable to handle the system dynamics directly in continuous time.

However, very few techniques exist for continuous-time RL and they lack flexibility in value function approximation.

In this paper, we propose a novel framework for model-based continuous-time value function approximation in reproducing kernel Hilbert spaces.

The resulting framework is so flexible that it can accommodate any kind of kernel-based approach, such as Gaussian processes and kernel adaptive filters, and it allows us to handle uncertainties and nonstationarity without prior knowledge about the environment or what basis functions to employ.

We demonstrate the validity of the presented framework through experiments.

Hybrid Knowledge Routed Modules for Large-scale Object Detection

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 coherence 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 Visual Genome (1000 categories) and 30.4% on ADE in terms of mAP.

Supervising Unsupervised Learning

Vikas Garg, Adam T. 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

Overlapping Clustering Models, and One (class) SVM to Bind Them All

Xueyu Mao, Purnamrita Sarkar, Deepayan Chakrabarti

People belong to multiple communities, words belong to multiple topics, and books cover multiple genres; overlapping clusters are commonplace. Many existing overlapping clustering methods model each person (or word, or book) as a non-negative weighted combination of "exemplars" who belong solely to one community, with some small noise. Geometrically, each person is a point on a cone whose corners are these exemplars. This basic form encompasses the widely used Mixed Membership Stochastic Blockmodel of networks and its degree-corrected variants, as well as topic models such as LDA. We show that a simple one-class SVM yields provably consistent parameter inference for all such models, and scales to large datasets. Experimental results on several simulated and real datasets show our algorithm (called SVM-cone) is both accurate and scalable.

BRITS: Bidirectional Recurrent Imputation for Time Series

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.

Improving Online Algorithms via ML Predictions

Manish Purohit, Zoya Svitkina, Ravi Kumar

In this work we study the problem of using machine-learned predictions to improve performance of online algorithms. We consider two classical problems, ski rental and non-clairvoyant job scheduling, and obtain new online algorithms that use predictions to make their decisions. These algorithms are oblivious to the performance of the predictor, improve with better predictions, but do not degrade much if the predictions are poor.

Learning Latent Subspaces in Variational Autoencoders

Jack Klys, Jake Snell, Richard Zemel

Variational autoencoders (VAEs) are widely used deep generative models capable of learning unsupervised latent representations of data. Such representations are often difficult to interpret or control. We consider the problem of unsupervised learning of features correlated to specific labels in a dataset. We propose a VAE-based generative model which we show is capable of extracting features correlated to binary labels in the data and structuring it in a latent subspace which is easy to interpret. Our model, the Conditional Subspace VAE (CSVAE), uses mutual information minimization to learn a low-dimensional latent subspace associated with each label that can easily be inspected and independently manipulated. We demonstrate the utility of the learned representations for attribute manipulation tasks on both the Toronto Face and CelebA datasets.

VideoCapsuleNet: A Simplified Network for Action Detection

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.

Causal Inference via Kernel Deviance Measures

Jovana Mitrovic, Dino Sejdinovic, Yee Whye Teh

Discovering the causal structure among a set of variables is a fundamental problem in many areas of science. In this paper, we propose Kernel Conditional Deviance for Causal Inference (KCDC) a fully nonparametric causal discovery method based on purely observational data. From a novel interpretation of the notion of asymmetry between cause and effect, we derive a corresponding asymmetry measure using the framework of reproducing kernel Hilbert spaces. Based on this, we propose three decision rules for causal discovery. We demonstrate the wide applicability and robustness of our method across a range of diverse synthetic datasets. Furthermore, we test our method on real-world time series data and the real-world benchmark dataset Tübingen Cause-Effect Pairs where we outperform state-of-the-art approaches.

Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects

Adam Kosior, Hyunjik Kim, Yee Whye Teh, Ingmar Posner

We present Sequential Attend, Infer, Repeat (SQAIR), an interpretable deep generative model for image sequences.

It can reliably discover and track objects through the sequence; it can also conditionally generate future frames, thereby simulating expected motion of objects.

This is achieved by explicitly encoding object numbers, locations and appearances in the latent variables of the model.

SQAIR retains all strengths of its predecessor, Attend, Infer, Repeat (AIR, Eslami et. al. 2016), including unsupervised learning, made possible by inductive biases present in the model structure.

We use a moving multi-\textsc{mnist} dataset to show limitations of AIR in detecting overlapping or partially occluded objects, and show how \textsc{sqair} overcomes them by leveraging temporal consistency of objects.

Finally, we also apply SQAIR to real-world pedestrian CCTV data, where it learns to reliably detect, track and generate walking pedestrians with no supervision.

Learning with SGD and Random Features

Luigi Carratino, Alessandro Rudi, Lorenzo Rosasco

Sketching and stochastic gradient methods are arguably the most common techniques to derive efficient large scale learning algorithms. In this paper, we investigate their application in the context of nonparametric statistical learning. More precisely, we study the estimator defined by stochastic gradient with mini batches and random features. The latter can be seen as form of nonlinear sketching and used to define approximate kernel methods. The considered estimator is not explicitly penalized/constrained and regularization is implicit. Indeed, our study highlights how different parameters, such as number of features, iterations, step-size and mini-batch size control the learning properties of the solution. We do this by deriving optimal finite sample bounds, under standard assumptions. The obtained results are corroborated and illustrated by numerical experiments.

Boolean Decision Rules via Column Generation

Sanjeeb Dash, Oktay Gunluk, Dennis Wei

This paper considers the learning of Boolean rules in either disjunctive normal form (DNF, OR-of-ANDs, equivalent to decision rule sets) or conjunctive normal form (CNF, AND-of-ORs) as an interpretable model for classification. An integer program is formulated to optimally trade classification accuracy for rule simplicity. Column generation (CG) is used to efficiently search over an exponential number of candidate clauses (conjunctions or disjunctions) without the need for heuristic rule mining. This approach also bounds the gap between the selected rule set and the best possible rule set on the training data. To handle large datasets, we propose an approximate CG algorithm using randomization. Compared to three recently proposed alternatives, the CG algorithm dominates the accuracy-simplicity trade-off in 8 out of 16 datasets. When maximized for accuracy, CG is competitive with rule learners designed for this purpose, sometimes finding significantly simpler solutions that are no less accurate.

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

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

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Boosting Black Box Variational Inference

Francesco Locatello, Gideon Dresdner, Rajiv Khanna, Isabel Valera, Gunnar Raetsch

Approximating a probability density in a tractable manner is a central task in Bayesian statistics. Variational Inference (VI) is a popular technique that achieves tractability by choosing a relatively simple variational approximation. Borrowing ideas from the classic boosting framework, recent approaches attempt to \emph{boost} VI by replacing the selection of a single density with an iteratively constructed mixture of densities. In order to guarantee convergence, previous works impose stringent assumptions that require significant effort for practitioners. Specifically, they require a custom implementation of the greedy step (called the LMO) for every probabilistic model with respect to an unnatural variational family of truncated distributions. Our work fixes these issues with novel theoretical and algorithmic insights. On the theoretical side, we show that boosting VI satisfies a relaxed smoothness assumption which is sufficient for the convergence of the functional Frank-Wolfe (FW) algorithm. Furthermore, we rephrase the LMO problem and propose to maximize the Residual ELBO (RELBO) which replaces the standard ELBO optimization in VI. These theoretical enhancements allow for black box implementation of the boosting subroutine. Finally, we present a stopping criterion drawn from the duality gap in the classic FW analyses and exhaustive experiments to illustrate the usefulness of our theoretical and algorithmic con

tributions.

Transfer of Deep Reactive Policies for MDP Planning

Aniket (Nick) Bajpai, Sankalp Garg, Mausam

Domain-independent probabilistic planners input an MDP description in a factored representation language such as PPDDL or RDDL, and exploit the specifics of the representation for faster planning. Traditional algorithms operate on each problem instance independently, and good methods for transferring experience from policies of other instances of a domain to a new instance do not exist. Recently, researchers have begun exploring the use of deep reactive policies, trained via deep reinforcement learning (RL), for MDP planning domains. One advantage of deep reactive policies is that they are more amenable to transfer learning.

Variational Bayesian Monte Carlo

Luigi Acerbi

Many probabilistic models of interest in scientific computing and machine learning have expensive, black-box likelihoods that prevent the application of standard techniques for Bayesian inference, such as MCMC, which would require access to the gradient or a large number of likelihood evaluations.

We introduce here a novel sample-efficient inference framework, Variational Bayesian Monte Carlo (VBMC). VBMC combines variational inference with Gaussian-process based, active-sampling Bayesian quadrature, using the latter to efficiently approximate the intractable integral in the variational objective.

Our method produces both a nonparametric approximation of the posterior distribution and an approximate lower bound of the model evidence, useful for model selection.

We demonstrate VBMC both on several synthetic likelihoods and on a neuronal model with data from real neurons. Across all tested problems and dimensions (up to $D = 10$), VBMC performs consistently well in reconstructing the posterior and the model evidence with a limited budget of likelihood evaluations, unlike other methods that work only in very low dimensions. Our framework shows great promise as a novel tool for posterior and model inference with expensive, black-box likelihoods.

Tangent: Automatic differentiation using source-code transformation for dynamically typed array programming

Bart van Merriënboer, Dan Moldovan, Alexander Wiltschko

The need to efficiently calculate first- and higher-order derivatives of increasingly complex models expressed in Python has stressed or exceeded the capabilities of available tools. In this work, we explore techniques from the field of automatic differentiation (AD) that can give researchers expressive power, performance and strong usability. These include source-code transformation (SCT), flexible gradient surgery, efficient in-place array operations, and higher-order derivatives. We implement and demonstrate these ideas in the Tangent software library for Python, the first AD framework for a dynamic language that uses SCT.

Learning Task Specifications from Demonstrations

Marcell Vazquez-Chanlatte, Susmit Jha, Ashish Tiwari, Mark K. Ho, Sanjit Seshia

Real-world applications often naturally decompose into several

sub-tasks. In many settings (e.g., robotics) demonstrations provide a natural way to specify the sub-tasks. However, most methods for learning from demonstrations either do not provide guarantees that the artifacts learned for the sub-tasks can be safely recombined or limit the types of composition available. Motivated by this deficit, we consider the problem of inferring Boolean non-Markovian rewards (also known as logical trace properties or specifications) from demonstrations provided by an agent operating in an uncertain, stochastic environment. Crucially, specifications admit well-defined composition rules that are typically easy to interpret. In this paper, we formulate the

specification inference task as a maximum a posteriori (MAP) probability inference problem, apply the principle of maximum entropy to derive an analytic demonstration likelihood model and give an efficient approach to search for the most likely specification in a large candidate pool of specifications. In our experiments, we demonstrate how learning specifications can help avoid common problems that often arise due to ad-hoc reward composition.

Sparse PCA from Sparse Linear Regression

Guy Bresler, Sung Min Park, Madalina Persu

Sparse Principal Component Analysis (SPCA) and Sparse Linear Regression (SLR) have a wide range of applications and have attracted a tremendous amount of attention in the last two decades as canonical examples of statistical problems in high dimension. A variety of algorithms have been proposed for both SPCA and SLR, but an explicit connection between the two had not been made. We show how to efficiently transform a black-box solver for SLR into an algorithm for SPCA: assuming the SLR solver satisfies prediction error guarantees achieved by existing efficient algorithms such as those based on the Lasso, the SPCA algorithm derived from it achieves near state of the art guarantees for testing and for support recovery for the single spiked covariance model as obtained by the current best polynomial-time algorithms. Our reduction not only highlights the inherent similarity between the two problems, but also, from a practical standpoint, allows one to obtain a collection of algorithms for SPCA directly from known algorithms for SLR. We provide experimental results on simulated data comparing our proposed framework to other algorithms for SPCA.

GILBO: One Metric to Measure Them All

Alexander A. Alemi, Ian Fischer

We propose a simple, tractable lower bound on the mutual information contained in the joint generative density of any latent variable generative model: the GILBO (Generative Information Lower Bound). It offers a data-independent measure of the complexity of the learned latent variable description, giving the log of the effective description length. It is well-defined for both VAEs and GANs. We compute the GILBO for 800 GANs and VAEs each trained on four datasets (MNIST, FashionMNIST, CIFAR-10 and CelebA) and discuss the results.

Maximizing Induced Cardinality Under a Determinantal Point Process

Jennifer A. Gillenwater, Alex Kulesza, Sergei Vassilvitskii, Zelda E. Mariet

Determinantal point processes (DPPs) are well-suited to recommender systems where the goal is to generate collections of diverse, high-quality items. In the existing literature this is usually formulated as finding the mode of the DPP (the so-called MAP set). However, the MAP objective inherently assumes that the DPP models "optimal" recommendation sets, and yet obtaining such a DPP is nontrivial when there is no ready source of example optimal sets. In this paper we advocate an alternative framework for applying DPPs to recommender systems. Our approach assumes that the DPP simply models user engagements with recommended items, which is more consistent with how DPPs for recommender systems are typically trained. With this assumption, we are able to formulate a metric that measures the expected number of items that a user will engage with. We formalize this optimization of this metric as the Maximum Induced Cardinality (MIC) problem. Although the MIC objective is not submodular, we show that it can be approximated by a submodular function, and that empirically it is well-optimized by a greedy algorithm.

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

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

her 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>.

Simple random search of static linear policies is competitive for reinforcement learning

Horia Mania, Aurelia Guy, Benjamin Recht

Model-free reinforcement learning aims to offer off-the-shelf solutions for controlling dynamical systems without requiring models of the system dynamics. We introduce a model-free random search algorithm for training static, linear policies for continuous control problems. Common evaluation methodology shows that our method matches state-of-the-art sample efficiency on the benchmark MuJoCo locomotion tasks. Nonetheless, more rigorous evaluation reveals that the assessment of performance on these benchmarks is optimistic. We evaluate the performance of our method over hundreds of random seeds and many different hyperparameter configurations for each benchmark task. This extensive evaluation is possible because of the small computational footprint of our method. Our simulations highlight a high variability in performance in these benchmark tasks, indicating that commonly used estimations of sample efficiency do not adequately evaluate the performance of RL algorithms. Our results stress the need for new baselines, benchmarks and evaluation methodology for RL algorithms.

Automatic differentiation in ML: Where we are and where we should be going

Bart van Merriënboer, Olivier Breuleux, Arnaud Bergeron, Pascal Lamblin

We review the current state of automatic differentiation (AD) for array programming in machine learning (ML), including the different approaches such as operator overloading (OO) and source transformation (ST) used for AD, graph-based intermediate representations for programs, and source languages. Based on these insights, we introduce a new graph-based intermediate representation (IR) which specifically aims to efficiently support fully-general AD for array programming. Unlike existing dataflow programming representations in ML frameworks, our IR naturally supports function calls, higher-order functions and recursion, making ML models easier to implement. The ability to represent closures allows us to perform AD using ST without a tape, making the resulting derivative (adjoint) program amenable to ahead-of-time optimization using tools from functional language compilers, and enabling higher-order derivatives. Lastly, we introduce a proof of concept compiler toolchain called Myia which uses a subset of Python as a front end.

Improving Neural Program Synthesis with Inferred Execution Traces

Eui Chul Shin, Illia Polosukhin, Dawn Song

The task of program synthesis, or automatically generating programs that are consistent with a provided specification, remains a challenging task in artificial intelligence. As in other fields of AI, deep learning-based end-to-end approaches have made great advances in program synthesis. However, more so than other fields such as computer vision, program synthesis provides greater opportunities to explicitly exploit structured information such as execution traces, which contain a superset of the information input/output pairs. While they are highly useful for program synthesis, as execution traces are more difficult to obtain than input/output pairs, we use the insight that we can split the process into two parts: infer the trace from the input/output example, then infer the program from t

he trace. This simple modification leads to state-of-the-art results in program synthesis in the Karel domain, improving accuracy to 81.3% from the 77.12% of prior work.

Discretely Relaxing Continuous Variables for tractable Variational Inference
Trefor Evans, Prasanth Nair

We explore a new research direction in Bayesian variational inference with discrete latent variable priors where we exploit Kronecker matrix algebra for efficient and exact computations of the evidence lower bound (ELBO). The proposed "DIRECT" approach has several advantages over its predecessors; (i) it can exactly compute ELBO gradients (i.e. unbiased, zero-variance gradient estimates), eliminating the need for high-variance stochastic gradient estimators and enabling the use of quasi-Newton optimization methods; (ii) its training complexity is independent of the number of training points, permitting inference on large datasets; and (iii) its posterior samples consist of sparse and low-precision quantized integers which permit fast inference on hardware limited devices. In addition, our DIRECT models can exactly compute statistical moments of the parameterized predictive posterior without relying on Monte Carlo sampling. The DIRECT approach is not practical for all likelihoods, however, we identify a popular model structure which is practical, and demonstrate accurate inference using latent variables discretized as extremely low-precision 4-bit quantized integers. While the ELBO computations considered in the numerical studies require over 10^{2352} log-likelihood evaluations, we train on datasets with over two-million points in just seconds.

The Limits of Post-Selection Generalization

Jonathan Ullman, Adam Smith, Kobbi Nissim, Uri Stemmer, Thomas Steinke

While statistics and machine learning offers numerous methods for ensuring generalization, these methods often fail in the presence of post selection---the common practice in which the choice of analysis depends on previous interactions with the same dataset. A recent line of work has introduced powerful, general purpose algorithms that ensure a property called post hoc generalization (Cummings et al., COLT'16), which says that no person when given the output of the algorithm should be able to find any statistic for which the data differs significantly from the population it came from.

Revisiting (ϵ, γ, τ) -similarity learning for domain adaptation
Sofiane Dhouib, Ievgen Redko

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning and Testing Causal Models with Interventions

Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, Saravanan Kandasamy

We consider testing and learning problems on causal Bayesian networks as defined by Pearl (Pearl, 2009). Given a causal Bayesian network M on a graph with n discrete variables and bounded in-degree and bounded ``confounded components'', we show that $O(\log n)$ interventions on an unknown causal Bayesian network X on the same graph, and $O(n/\epsilon^2)$ samples per intervention, suffice to efficiently distinguish whether $X=M$ or whether there exists some intervention under which X and M are farther than ϵ in total variation distance. We also obtain sample/time/intervention efficient algorithms for: (i) testing the identity of two unknown causal Bayesian networks on the same graph; and (ii) learning a causal Bayesian network on a given graph. Although our algorithms are non-adaptive, we show that adaptivity does not help in general: $\Omega(\log n)$ interventions are necessary for testing the identity of two unknown causal Bayesian networks on the same graph, even adaptively. Our algorithms are enabled by a new subadditivity inequality for the squared Hellinger distance between two causal Bayesian networks.

S.

Evolved Policy Gradients

Rein Houthoofd, 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.

Demystifying excessively volatile human learning: A Bayesian persistent prior and a neural approximation

Chaitanya Ryali, Gautam Reddy, Angela J. Yu

Understanding how humans and animals learn about statistical regularities in stable and volatile environments, and utilize these regularities to make predictions and decisions, is an important problem in neuroscience and psychology. Using a Bayesian modeling framework, specifically the Dynamic Belief Model (DBM), it has previously been shown that humans tend to make the {\it default} assumption that environmental statistics undergo abrupt, unsignaled changes, even when environmental statistics are actually stable. Because exact Bayesian inference in this setting, an example of switching state space models, is computationally intense, a number of approximately Bayesian and heuristic algorithms have been proposed to account for learning/prediction in the brain. Here, we examine a neurally plausible algorithm, a special case of leaky integration dynamics we denote as EXP (for exponential filtering), that is significantly simpler than all previously suggested algorithms except for the delta-learning rule, and which far outperforms the delta rule in approximating Bayesian prediction performance. We derive the theoretical relationship between DBM and EXP, and show that EXP gains computational efficiency by foregoing the representation of inferential uncertainty (as does the delta rule), but that it nevertheless achieves near-Bayesian performance due to its ability to incorporate a "persistent prior" influence unique to DBM and absent from the other algorithms. Furthermore, we show that EXP is comparable to DBM but better than all other models in reproducing human behavior in a visual search task, suggesting that human learning and prediction also incorporate an element of persistent prior. More broadly, our work demonstrates that when observations are information-poor, detecting changes or modulating the learning rate is both {\it difficult} and (thus) {\it unnecessary} for making Bayes-optimal predictions.

Distributionally Robust Graphical Models

Rizal Fathony, Ashkan Rezaei, Mohammad Ali Bashiri, Xinhua Zhang, Brian Ziebart

In many structured prediction problems, complex relationships between variables are compactly defined using graphical structures. The most prevalent graphical prediction methods---probabilistic graphical models and large margin methods---have their own distinct strengths but also possess significant drawbacks. Conditional random fields (CRFs) are Fisher consistent, but they do not permit integration of customized loss metrics into their learning process. Large-margin models, such as structured support vector machines (SSVMs), have the flexibility to incorporate customized loss metrics, but lack Fisher consistency guarantees. We present adversarial graphical models (AGM), a distributionally robust approach for constructing a predictor that performs robustly for a class of data distributions defined using a graphical structure. Our approach enjoys both the flexibility of incorporating customized loss metrics into its design as well as the statisti

cal guarantee of Fisher consistency. We present exact learning and prediction algorithms for AGM with time complexity similar to existing graphical models and show the practical benefits of our approach with experiments.

Natasha 2: Faster Non-Convex Optimization Than SGD

Zeyuan Allen-Zhu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Iterative Value-Aware Model Learning

Amir-massoud Farahmand

This paper introduces a model-based reinforcement learning (MBRL) framework that incorporates the underlying decision problem in learning the transition model of the environment. This is in contrast with conventional approaches to MBRL that learn the model of the environment, for example by finding the maximum likelihood estimate, without taking into account the decision problem. Value-Aware Model Learning (VAML) framework argues that this might not be a good idea, especially if the true model of the environment does not belong to the model class from which we are estimating the model. The original VAML framework, however, may result in an optimization problem that is difficult to solve. This paper introduces a new MBRL class of algorithms, called Iterative VAML, that benefits from the structure of how the planning is performed (i.e., through approximate value iteration) to devise a simpler optimization problem. The paper theoretically analyzes Iterative VAML and provides finite sample error upper bound guarantee for it.

PCA of high dimensional random walks with comparison to neural network training

Joseph Antognini, Jascha Sohl-Dickstein

One technique to visualize the training of neural networks is to perform PCA on the parameters over the course of training and to project to the subspace spanned by the first few PCA components. In this paper we compare this technique to the PCA of a high dimensional random walk. We compute the eigenvalues and eigenvectors of the covariance of the trajectory and prove that in the long trajectory and high dimensional limit most of the variance is in the first few PCA components, and that the projection of the trajectory onto any subspace spanned by PCA components is a Lissajous curve. We generalize these results to a random walk with momentum and to an Ornstein-Uhlenbeck processes (i.e., a random walk in a quadratic potential) and show that in high dimensions the walk is not mean reverting, but will instead be trapped at a fixed distance from the minimum. We finally analyze PCA projected training trajectories for: a linear model trained on CIFAR-10; a fully connected model trained on MNIST; and ResNet-50-v2 trained on ImageNet. In all cases, both the distribution of PCA eigenvalues and the projected trajectories resemble those of a random walk with drift.

Learning Libraries of Subroutines for Neurally-Guided Bayesian Program Induction

Kevin Ellis, Lucas Morales, Mathias Sablé-Meyer, Armando Solar-Lezama, Josh Tenenbaum

Successful approaches to program induction require a hand-engineered domain-specific language (DSL), constraining the space of allowed programs and imparting prior knowledge of the domain. We contribute a program induction algorithm that learns a DSL while jointly training a neural network to efficiently search for programs in the learned DSL. We use our model to synthesize functions on lists, edit text, and solve symbolic regression problems, showing how the model learns a domain-specific library of program components for expressing solutions to problems in the domain.

Streaming Kernel PCA with $\tilde{O}(\sqrt{n})$ Random Features

Enayat Ullah, Poorya Mianjy, Teodor Vanislavov Marinov, Raman Arora

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Faster Neural Networks Straight from JPEG

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.

Bayesian Model Selection Approach to Boundary Detection with Non-Local Priors

Fei Jiang, Guosheng Yin, Francesca Dominici

Based on non-local prior distributions, we propose a Bayesian model selection (BMS) procedure for boundary detection in a sequence of data with multiple systematic mean changes. The BMS method can effectively suppress the non-boundary spike points with large instantaneous changes. We speed up the algorithm by reducing the multiple change points to a series of single change point detection problems. We establish the consistency of the estimated number and locations of the change points under various prior distributions. Extensive simulation studies are conducted to compare the BMS with existing methods, and our approach is illustrated with application to the magnetic resonance imaging guided radiation therapy data.

Densely Connected Attention Propagation for Reading Comprehension

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.

Query Complexity of Bayesian Private Learning

Kuang Xu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

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

Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling

Emilie Kaufmann, Wouter M. Koolen, Aurélien Garivier

Learning the minimum/maximum mean among a finite set of distributions is a fundamental sub-problem in planning, game tree search and reinforcement learning. We formalize this learning task as the problem of sequentially testing how the minimum mean among a finite set of distributions compares to a given threshold. We develop refined non-asymptotic lower bounds, which show that optimality mandates very different sampling behavior for a low vs high true minimum. We show that Thompson Sampling and the intuitive Lower Confidence Bounds policy each nail only one of these cases. We develop a novel approach that we call Murphy Sampling. Even though it entertains exclusively low true minima, we prove that MS is optimal for both possibilities. We then design advanced self-normalized deviation inequalities, fueling more aggressive stopping rules. We complement our theoretical guarantees by experiments showing that MS works best in practice.

Content preserving text generation with attribute controls

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.

Latent Gaussian Activity Propagation: Using Smoothness and Structure to Separate and Localize Sounds in Large Noisy Environments

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.

Differentially Private Testing of Identity and Closeness of Discrete Distributions

Jayadev Acharya, Ziteng Sun, Huanyu Zhang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Non-Ergodic Alternating Proximal Augmented Lagrangian Algorithms with Optimal Rates

Quoc Tran Dinh

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Multi-objective Maximization of Monotone Submodular Functions with Cardinality Constraint

Rajan Udwani

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Fully Understanding The Hashing Trick

Casper B. Freksen, Lior Kamma, Kasper Green Larsen

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Policy-Conditioned Uncertainty Sets for Robust Markov Decision Processes

Andrea Tirinzoni, Marek Petrik, Xiangli Chen, Brian Ziebart

What policy should be employed in a Markov decision process with uncertain parameters? Robust optimization answer to this question is to use rectangular uncertainty sets, which independently reflect available knowledge about each state, and then obtains a decision policy that maximizes expected reward for the worst-case decision process parameters from these uncertainty sets. While this rectangularity is convenient computationally and leads to tractable solutions, it often produces policies that are too conservative in practice, and does not facilitate knowledge transfer between portions of the state space or across related decision processes. In this work, we propose non-rectangular uncertainty sets that bound marginal moments of state-action features defined over entire trajectories through a decision process. This enables generalization to different portions of the state space while retaining appropriate uncertainty of the decision process. We develop algorithms for solving the resulting robust decision problems, which reduce to finding an optimal policy for a mixture of decision processes, and demonstrate the benefits of our approach experimentally.

Visual Reinforcement Learning with Imagined Goals

Ashvin V. Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, Sergey Levine

For an autonomous agent to fulfill a wide range of user-specified goals at test time, it must be able to learn broadly applicable and general-purpose skill repertoires. Furthermore, to provide the requisite level of generality, these skills must handle raw sensory input such as images. In this paper, we propose an algorithm that acquires such general-purpose skills by combining unsupervised representation learning and reinforcement learning of goal-conditioned policies. Since the particular goals that might be required at test-time are not known in advance, the agent performs a self-supervised "practice" phase where it imagines goal

s and attempts to achieve them. We learn a visual representation with three distinct purposes: sampling goals for self-supervised practice, providing a structured transformation of raw sensory inputs, and computing a reward signal for goal reaching. We also propose a retroactive goal relabeling scheme to further improve the sample-efficiency of our method. Our off-policy algorithm is efficient enough to learn policies that operate on raw image observations and goals in a real-world physical system, and substantially outperforms prior techniques.

DropBlock: A regularization method for convolutional networks

Golnaz Ghiasi, Tsung-Yi Lin, Quoc V. Le

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

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

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.

RetGK: Graph Kernels based on Return Probabilities of Random Walks

Zhen Zhang, Mianzhi Wang, Yijian Xiang, Yan Huang, Arye Nehorai

Graph-structured data arise in wide applications, such as computer vision, bioinformatics, and social networks. Quantifying similarities among graphs is a fundamental problem. In this paper, we develop a framework for computing graph kernels, based on return probabilities of random walks. The advantages of our proposed kernels are that they can effectively exploit various node attributes, while being scalable to large datasets. We conduct extensive graph classification experiments to evaluate our graph kernels. The experimental results show that our graph kernels significantly outperform other state-of-the-art approaches in both accuracy and computational efficiency.

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

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 dataset by a large margin, and is also able to handle large-scale heterogeneous inputs (real and synthetic images) with missing annotation modalities.

Gradient Descent Meets Shift-and-Invert Preconditioning for Eigenvector Computation

Zhiqiang Xu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Inequity aversion improves cooperation in intertemporal social dilemmas

Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, Thore Graepel

Groups of humans are often able to find ways to cooperate with one another in complex, temporally extended social dilemmas. Models based on behavioral economics are only able to explain this phenomenon for unrealistic stateless matrix games. Recently, multi-agent reinforcement learning has been applied to generalize social dilemma problems to temporally and spatially extended Markov games. However, this has not yet generated an agent that learns to cooperate in social dilemmas as humans do. A key insight is that many, but not all, human individuals have inequity averse social preferences. This promotes a particular resolution of the matrix game social dilemma wherein inequity-averse individuals are personally pro-social and punish defectors. Here we extend this idea to Markov games and show that it promotes cooperation in several types of sequential social dilemma, via a profitable interaction with policy learnability. In particular, we find that inequity aversion improves temporal credit assignment for the important class of intertemporal social dilemmas. These results help explain how large-scale cooperation may emerge and persist.

Towards Deep Conversational Recommendations

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, Chris Pal

There has been growing interest in using neural networks and deep learning techniques to create dialogue systems. Conversational recommendation is an interesting setting for the scientific exploration of dialogue with natural language as the associated discourse involves goal-driven dialogue that often transforms naturally into more free-form chat. This paper provides two contributions. First, until now there has been no publicly available large-scale data set consisting of real-world dialogues centered around recommendations.

To address this issue and to facilitate our exploration here, we have collected ReDial, a data set consisting of over 10,000 conversations centered around the theme of providing movie recommendations. We make this data available to the community for further research. Second, we use this dataset to explore multiple facets of conversational recommendations. In particular we explore new neural architectures, mechanisms and methods suitable for composing conversational recommendation systems. Our dataset allows us to systematically probe model sub-components addressing different parts of the overall problem domain ranging from: sentiment analysis and cold-start recommendation generation to detailed aspects of how natural language is used in this setting in the real world. We combine such sub-components into a full-blown dialogue system and examine its behavior.

Deep Generative Models for Distribution-Preserving Lossy Compression

Michael Tschannen, Eiríkur Agustsson, Mario Lucic

We propose and study the problem of distribution-preserving lossy compression. M

otivated by recent advances in extreme image compression which allow to maintain artifact-free reconstructions even at very low bitrates, we propose to optimize the rate-distortion tradeoff under the constraint that the reconstructed samples follow the distribution of the training data. The resulting compression system recovers both ends of the spectrum: On one hand, at zero bitrate it learns a generative model of the data, and at high enough bitrates it achieves perfect reconstruction. Furthermore, for intermediate bitrates it smoothly interpolates between learning a generative model of the training data and perfectly reconstructing the training samples. We study several methods to approximately solve the proposed optimization problem, including a novel combination of Wasserstein GAN and Wasserstein Autoencoder, and present an extensive theoretical and empirical characterization of the proposed compression systems.

With Friends Like These, Who Needs Adversaries?

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.

Learning to Teach with Dynamic Loss Functions

Lijun Wu, Fei Tian, Yingce Xia, Yang Fan, Tao Qin, Lai Jian-Huang, Tie-Yan Liu

Teaching is critical to human society: it is with teaching that prospective students are educated and human civilization can be inherited and advanced. A good teacher not only provides his/her students with qualified teaching materials (e.g., textbooks), but also sets up appropriate learning objectives (e.g., course projects and exams) considering different situations of a student. When it comes to artificial intelligence, treating machine learning models as students, the loss functions that are optimized act as perfect counterparts of the learning objective set by the teacher. In this work, we explore the possibility of imitating human teaching behaviors by dynamically and automatically outputting appropriate loss functions to train machine learning models. Different from typical learning settings in which the loss function of a machine learning model is predefined and fixed, in our framework, the loss function of a machine learning model (we call it student) is defined by another machine learning model (we call it teacher). The ultimate goal of teacher model is cultivating the student to have better performance measured on development dataset. Towards that end, similar to human teaching, the teacher, a parametric model, dynamically outputs different loss functions that will be used and optimized by its student model at different training stages. We develop an efficient learning method for the teacher model that makes gradient based optimization possible, exempt of the ineffective solutions such as policy optimization. We name our method as 'learning to teach with dynamic loss functions' (L2T-DLF for short). Extensive experiments on real world tasks including image classification and neural machine translation demonstrate that our method significantly improves the quality of various student models.

Learning Bounds for Greedy Approximation with Explicit Feature Maps from Multiple Kernels

Shahin Shahrampour, Vahid Tarokh

Nonlinear kernels can be approximated using finite-dimensional feature maps for efficient risk minimization. Due to the inherent trade-off between the dimension of the (mapped) feature space and the approximation accuracy, the key problem is to identify promising (explicit) features leading to a satisfactory out-of-sample performance. In this work, we tackle this problem by efficiently choosing such features from multiple kernels in a greedy fashion. Our method sequentially selects these explicit features from a set of candidate features using a correlation metric. We establish an out-of-sample error bound capturing the trade-off between the error in terms of explicit features (approximation error) and the error due to spectral properties of the best model in the Hilbert space associated to the combined kernel (spectral error). The result verifies that when the (best) underlying data model is sparse enough, i.e., the spectral error is negligible, one can control the test error with a small number of explicit features, that can scale poly-logarithmically with data. Our empirical results show that given a fixed number of explicit features, the method can achieve a lower test error with a smaller time cost, compared to the state-of-the-art in data-dependent random features.

Distributed Multitask Reinforcement Learning with Quadratic Convergence

Rasul Tutunov, Dongho Kim, Haitham Bou Ammar

Multitask reinforcement learning (MTRL) suffers from scalability issues when the number of tasks or trajectories grows large. The main reason behind this drawback is the reliance on centralized solutions. Recent methods exploited the connection between MTRL and general consensus to propose scalable solutions. These methods, however, suffer from two drawbacks. First, they rely on predefined objectives, and, second, exhibit linear convergence guarantees. In this paper, we improve over state-of-the-art by deriving multitask reinforcement learning from a variational inference perspective. We then propose a novel distributed solver for MTRL with quadratic convergence guarantees.

The Global Anchor Method for Quantifying Linguistic Shifts and Domain Adaptation

Zi Yin, Vin Sachidananda, Balaji Prabhakar

Language is dynamic, constantly evolving and adapting with respect to time, domain or topic. The adaptability of language is an active research area, where researchers discover social, cultural and domain-specific changes in language using distributional tools such as word embeddings. In this paper, we introduce the global anchor method for detecting corpus-level language shifts. We show both theoretically and empirically that the global anchor method is equivalent to the alignment method, a widely-used method for comparing word embeddings, in terms of detecting corpus-level language shifts. Despite their equivalence in terms of detection abilities, we demonstrate that the global anchor method is superior in terms of applicability as it can compare embeddings of different dimensionalities.

Furthermore, the global anchor method has implementation and parallelization advantages. We show that the global anchor method reveals fine structures in the evolution of language and domain adaptation. When combined with the graph Laplacian technique, the global anchor method recovers the evolution trajectory and domain clustering of disparate text corpora.

Understanding Weight Normalized Deep Neural Networks with Rectified Linear Units

Yixi Xu, Xiao Wang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Low-shot Learning via Covariance-Preserving Adversarial Augmentation Networks

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.

Bilevel Distance Metric Learning for Robust Image Recognition

Jie Xu, Lei Luo, Cheng Deng, Heng Huang

Metric learning, aiming to learn a discriminative Mahalanobis distance matrix M that can effectively reflect the similarity between data samples, has been widely studied in various image recognition problems. Most of the existing metric learning methods input the features extracted directly from the original data in the preprocess phase. What's worse, these features usually take no consideration of the local geometrical structure of the data and the noise existed in the data, thus they may not be optimal for the subsequent metric learning task. In this paper, we integrate both feature extraction and metric learning into one joint optimization framework and propose a new bilevel distance metric learning model. Specifically, the lower level characterizes the intrinsic data structure using graph regularized sparse coefficients, while the upper level forces the data samples from the same class to be close to each other and pushes those from different classes far away.

In addition, leveraging the KKT conditions and the alternating direction method (ADM), we derive an efficient algorithm to solve the proposed new model. Extensive experiments on various occluded datasets demonstrate the effectiveness and robustness of our method.

Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models

Amir Dezfouli, Richard Morris, Fabio T. Ramos, Peter Dayan, Bernard Balleine

Neuroscience studies of human decision-making abilities commonly involve subjects completing a decision-making task while BOLD signals are recorded using fMRI. Hypotheses are tested about which brain regions mediate the effect of past experience, such as rewards, on future actions. One standard approach to this is model-based fMRI data analysis, in which a model is fitted to the behavioral data, i.e., a subject's choices, and then the neural data are parsed to find brain regions whose BOLD signals are related to the model's internal signals. However, the internal mechanics of such purely behavioral models are not constrained by the neural data, and therefore might miss or mischaracterize aspects of the brain. To address this limitation, we introduce a new method using recurrent neural network models that are flexible enough to be jointly fitted to the behavioral and neural data. We trained a model so that its internal states were suitably related to neural activity during the task, while at the same time its output predicted the next action a subject would execute. We then used the fitted model to create a novel visualization of the relationship between the activity in brain regions at different times following a reward and the choices the subject subsequently made. Finally, we validated our method using a previously published dataset. We found that the model was able to recover the underlying neural substrates that were

discovered by explicit model engineering in the previous work, and also derived new results regarding the temporal pattern of brain activity.

Learning from Group Comparisons: Exploiting Higher Order Interactions

Yao Li, Minhao Cheng, Kevin Fujii, Fushing Hsieh, Cho-Jui Hsieh

We study the problem of learning from group comparisons, with applications in predicting outcomes of sports and online games. Most of the previous works in this area focus on learning individual effects---they assume each player has an underlying score, and the 'ability' of the team is modeled by the sum of team members' scores. Therefore, all the current approaches cannot model deeper interaction between team members: some players perform much better if they play together, and some players perform poorly together. In this paper, we propose a new model that takes the player-interaction effects into consideration. However, under certain circumstances, the total number of individuals can be very large, and number of player interactions grows quadratically, which makes learning intractable.

In this case, we propose a latent factor model, and show that the sample complexity of our model is bounded under mild assumptions. Finally, we show that our proposed models have much better prediction power on several E-sports datasets, and furthermore can be used to reveal interesting patterns that cannot be discovered by previous methods.

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

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.

A Bandit Approach to Sequential Experimental Design with False Discovery Control

Kevin G. Jamieson, Lalit Jain

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

HitNet: Hybrid Ternary Recurrent Neural Network

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 differen

t 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.

SLAYER: Spike Layer Error Reassignment in Time

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.

A Convex Duality Framework for GANs

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.

Empirical Risk Minimization Under Fairness Constraints

Michele Donini, Luca Oneto, Shai Ben-David, John S. Shawe-Taylor, Massimiliano Pontil

We address the problem of algorithmic fairness: ensuring that sensitive informat

ion 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.

End-to-End Differentiable Physics for Learning and Control

Filipe de Avila Belbute-Peres, Kevin Smith, Kelsey Allen, Josh Tenenbaum, J. Zico Kolter

We present a differentiable physics engine that can be integrated as a module in deep neural networks for end-to-end learning. As a result, structured physics knowledge can be embedded into larger systems, allowing them, for example, to match observations by performing precise simulations, while achieving high sample efficiency. Specifically, in this paper we demonstrate how to perform backpropagation analytically through a physical simulator defined via a linear complementarity problem. Unlike traditional finite difference methods, such gradients can be computed analytically, which allows for greater flexibility of the engine. Through experiments in diverse domains, we highlight the system's ability to learn physical parameters from data, efficiently match and simulate observed visual behavior, and readily enable control via gradient-based planning methods. Code for the engine and experiments is included with the paper.

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

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.

Horizon-Independent Minimax Linear Regression

Alan Malek, Peter L. Bartlett

We consider online linear regression: at each round, an adversary reveals a covariate vector, the learner predicts a real value, the adversary reveals a label, and the learner suffers the squared prediction error. The aim is to minimize the difference between the cumulative loss and that of the linear predictor that is best in hindsight. Previous work demonstrated that the minimax optimal strategy is easy to compute recursively from the end of the game; this requires the entire sequence of covariate vectors in advance. We show that, once provided with a measure of the scale of the problem, we can invert the recursion and play the minimax strategy without knowing the future covariates. Further, we show that this forward recursion remains optimal even against adaptively chosen labels and covariates, provided that the adversary adheres to a set of constraints that prevent misrepresentation of the scale of the problem. This strategy is horizon-independent in that the regret and minimax strategies depend on the size of the constraint set and not on the time-horizon, and hence it incurs no more regret than the optimal strategy that knows in advance the number of rounds of the game. We also provide an interpretation of the minimax algorithm as a follow-the-regularized-leader strategy with a data-dependent regularizer and obtain an explicit expression for the minimax regret.

Support Recovery for Orthogonal Matching Pursuit: Upper and Lower bounds

Raghav Somani, Chirag Gupta, Prateek Jain, Praneeth Netrapalli

This paper studies the problem of sparse regression where the goal is to learn a sparse vector that best optimizes a given objective function. Under the assumption that the objective function satisfies restricted strong convexity (RSC), we analyze orthogonal matching pursuit (OMP), a greedy algorithm that is used heavily in applications, and obtain support recovery result as well as a tight generalization error bound for OMP. Furthermore, we obtain lower bounds for OMP, showing that both our results on support recovery and generalization error are tight up to logarithmic factors. To the best of our knowledge, these support recovery and generalization bounds are the first such matching upper and lower bounds (up to logarithmic factors) for $\{\text{any}\}$ sparse regression algorithm under the RSC assumption.

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

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 $\{\text{it statistical typicality}\}$ is high, in consonance with Barlow's seminaefficient 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.

The committee machine: Computational to statistical gaps in learning a two-layer neural network

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, Lenka Zdeborová

Heuristic tools from statistical physics have been used in the past to compute the optimal learning and generalization errors in the teacher-student scenario in multi-layer neural networks. In this contribution, we provide a rigorous justification of these approaches for a two-layers neural network model called the committee machine. We also introduce a version of the approximate message passing (AMP) algorithm for the committee machine that allows to perform optimal learning in polynomial time for a large set of parameters. We find that there are regimes in which a low generalization error is information-theoretically achievable while the AMP algorithm fails to deliver it; strongly suggesting that no efficient algorithm exists for those cases, and unveiling a large computational gap.

Adversarial vulnerability for any classifier

Alhussein Fawzi, Hamza Fawzi, Omar Fawzi

Despite achieving impressive performance, state-of-the-art classifiers remain highly vulnerable to small, imperceptible, adversarial perturbations. This vulnerability has proven empirically to be very intricate to address. In this paper, we study the phenomenon of adversarial perturbations under the assumption that the data is generated with a smooth generative model. We derive fundamental upper bounds on the robustness to perturbations of any classification function, and prove the existence of adversarial perturbations that transfer well across different classifiers with small risk. Our analysis of the robustness also provides insights onto key properties of generative models, such as their smoothness and dimensionality of latent space. We conclude with numerical experimental results showing that our bounds provide informative baselines to the maximal achievable robustness on several datasets.

A Deep Bayesian Policy Reuse Approach Against Non-Stationary Agents

YAN ZHENG, Zhaopeng Meng, Jianye Hao, Zongzhang Zhang, Tianpei Yang, Changjie Fan

In multiagent domains, coping with non-stationary agents that change behaviors from time to time is a challenging problem, where an agent is usually required to be able to quickly detect the other agent's policy during online interaction, and then adapt its own policy accordingly. This paper studies efficient policy detecting and reusing techniques when playing against non-stationary agents in Markov games. We propose a new deep BPR+ algorithm by extending the recent BPR+ algorithm with a neural network as the value-function approximator. To detect policy accurately, we propose the \textit{rectified belief model} taking advantage of the \textit{opponent model} to infer the other agent's policy from reward signals and its behaviors. Instead of directly storing individual policies as BPR+, we introduce \textit{distilled policy network} that serves as the policy library in BPR+, using policy distillation to achieve efficient online policy learning and reuse. Deep BPR+ inherits all the advantages of BPR+ and empirically shows better performance in terms of detection accuracy, cumulative rewards and speed of convergence compared to existing algorithms in complex Markov games with raw visual inputs.

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

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.

Trajectory Convolution for Action Recognition

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.

Adversarial Attacks on Stochastic Bandits

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, Jerry Zhu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Evolution-Guided Policy Gradient in Reinforcement Learning

Shauharda Khadka, Kagan Tumer

Deep Reinforcement Learning (DRL) algorithms have been successfully applied to a range of challenging control tasks. However, these methods typically suffer from three core difficulties: temporal credit assignment with sparse rewards, lack of effective exploration, and brittle convergence properties that are extremely sensitive to hyperparameters. Collectively, these challenges severely limit the applicability of these approaches to real world problems. Evolutionary Algorithms (EAs), a class of black box optimization techniques inspired by natural evolution, are well suited to address each of these three challenges. However, EAs typically suffer from high sample complexity and struggle to solve problems that require optimization of a large number of parameters. In this paper, we introduce Evolutionary Reinforcement Learning (ERL), a hybrid algorithm that leverages the population of an EA to provide diversified data to train an RL agent, and reinjects the RL agent into the EA population periodically to inject gradient information into the EA. ERL inherits EA's ability of temporal credit assignment with a fitness metric, effective exploration with a diverse set of policies, and stability of a population-based approach and complements it with off-policy DRL's ability to leverage gradients for higher sample efficiency and faster learning. Experiments in a range of challenging continuous control benchmarks demonstrate that ERL significantly outperforms prior DRL and EA methods.

Policy Regret in Repeated Games

Raman Arora, Michael Dinitz, Teodor Vanislavov Marinov, Mehryar Mohri

The notion of policy regret'' in online learning is supposed to capture the reactions of the adversary to the actions taken by the learner, which more traditional notions such as external regret do not take into account. We revisit this notion of policy regret, and first show that there are online learning settings in which policy regret and external regret are incompatible: any sequence of play which does well with respect to one must do poorly with respect to the other. We then focus on the game theoretic setting, when the adversary is a self-interested agent. In this setting we show that the external regret and policy regret are not in conflict, and in fact that a wide class of algorithms can ensure both as long as the adversary is also using such an algorithm. We also define a new notion of equilibrium which we call apolicy equilibrium'', and show that no-policy regret algorithms will have play which converges to such an equilibrium. Relating this back to external regret, we show that coarse correlated equilibria (which no-external regret players will converge to) are a strict subset of policy equilibria. So in game-theoretic settings every sequence of play with no external regret also has no policy regret, but the converse is not true.

Causal Inference with Noisy and Missing Covariates via Matrix Factorization

Nathan Kallus, Xiaojie Mao, Madeleine Udell

Valid causal inference in observational studies often requires controlling for confounders. However, in practice measurements of confounders may be noisy, and can lead to biased estimates of causal effects. We show that we can reduce bias induced by measurement noise using a large number of noisy measurements of the underlying confounders. We propose the use of matrix factorization to infer the co

founders from noisy covariates. This flexible and principled framework adapts to missing values, accommodates a wide variety of data types, and can enhance a wide variety of causal inference methods. We bound the error for the induced average treatment effect estimator and show it is consistent in a linear regression setting, using Exponential Family Matrix Completion preprocessing. We demonstrate the effectiveness of the proposed procedure in numerical experiments with both synthetic data and real clinical data.

Bayesian Distributed Stochastic Gradient Descent

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.

Benefits of over-parameterization with EM

Ji Xu, Daniel J. Hsu, Arian Maleki

Expectation Maximization (EM) is among the most popular algorithms for maximum likelihood estimation, but it is generally only guaranteed to find its stationary points of the log-likelihood objective. The goal of this article is to present theoretical and empirical evidence that over-parameterization can help EM avoid spurious local optima in the log-likelihood. We consider the problem of estimating the mean vectors of a Gaussian mixture model in a scenario where the mixing weights are known. Our study shows that the global behavior of EM, when one uses an over-parameterized model in which the mixing weights are treated as unknown, is better than that when one uses the (correct) model with the mixing weights fixed to the known values. For symmetric Gaussians mixtures with two components, we prove that introducing the (statistically redundant) weight parameters enables EM to find the global maximizer of the log-likelihood starting from almost any initial mean parameters, whereas EM without this over-parameterization may very often fail. For other Gaussian mixtures, we provide empirical evidence that shows similar behavior. Our results corroborate the value of over-parameterization in solving non-convex optimization problems, previously observed in other domains.

How to tell when a clustering is (approximately) correct using convex relaxations

Marina Meila

We introduce the Sublevel Set (SS) method, a generic method to obtain sufficient guarantees of near-optimality and uniqueness (up to small perturbations) for a clustering. This method can be instantiated for a variety of clustering loss functions for which convex relaxations exist. Obtaining the guarantees in practice amounts to solving a convex optimization. We demonstrate the applicability of this method by obtaining distribution free guarantees for K-means clustering on realistic data sets.

Faithful Inversion of Generative Models for Effective Amortized Inference

Stefan Webb, Adam Golinski, Rob Zinkov, Siddharth N, Tom Rainforth, Yee Whye Teh, Frank Wood

Inference amortization methods share information across multiple posterior-inference problems, allowing each to be carried out more efficiently. Generally, they require the inversion of the dependency structure in the generative model, as the modeller must learn a mapping from observations to distributions approximating the posterior. Previous approaches have involved inverting the dependency structure in a heuristic way that fails to capture these dependencies correctly, thereby limiting the achievable accuracy of the resulting approximations. We introduce an algorithm for faithfully, and minimally, inverting the graphical model structure of any generative model. Such inverses have two crucial properties: (a) they do not encode any independence assertions that are absent from the model and; (b) they are local maxima for the number of true independencies encoded. We prove the correctness of our approach and empirically show that the resulting minimally faithful inverses lead to better inference amortization than existing heuristic approaches.

TETRIS: Tile-matching the TREmendous Irregular Sparsity

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 attracted 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.

Stochastic Chebyshev Gradient Descent for Spectral Optimization

Insu Han, Haim Avron, Jinwoo Shin

A large class of machine learning techniques requires the solution of optimization problems involving spectral functions of parametric matrices, e.g. log-determinant and nuclear norm. Unfortunately, computing the gradient of a spectral function is generally of cubic complexity, as such gradient descent methods are rather expensive for optimizing objectives involving the spectral function. Thus, one naturally turns to stochastic gradient methods in hope that they will provide a way to reduce or altogether avoid the computation of full gradients. However, here a new challenge appears: there is no straightforward way to compute unbiased stochastic gradients for spectral functions. In this paper, we develop unbiased stochastic gradients for spectral-sums, an important subclass of spectral functions. Our unbiased stochastic gradients are based on combining randomized trace estimators with stochastic truncation of the Chebyshev expansions. A careful design of the truncation distribution allows us to offer distributions that are variance-optimal, which is crucial for fast and stable convergence of stochastic gradient methods. We further leverage our proposed stochastic gradients to devise stochastic methods for objective functions involving spectral-sums, and rigorously analyze their convergence rate. The utility of our methods is demonstrated in numerical experiments.

Model-based targeted dimensionality reduction for neuronal population data

Mikio Aoi, Jonathan W. Pillow

Summarizing high-dimensional data using a small number of parameters is a ubiquitous first step in the analysis of neuronal population activity. Recently developed methods use "targeted" approaches that work by identifying multiple, distinct low-dimensional subspaces of activity that capture the population response to individual experimental task variables, such as the value of a presented stimulus or the behavior of the animal. These methods have gained attention because they decompose total neural activity into what are ostensibly different parts of a neuronal computation. However, existing targeted methods have been developed outside of the confines of probabilistic modeling, making some aspects of the procedures ad hoc, or limited in flexibility or interpretability. Here we propose a new model-based method for targeted dimensionality reduction based on a probabilistic generative model of the population response data. The low-dimensional structure

cture of our model is expressed as a low-rank factorization of a linear regression model. We perform efficient inference using a combination of expectation maximization and direct maximization of the marginal likelihood. We also develop an efficient method for estimating the dimensionality of each subspace. We show that our approach outperforms alternative methods in both mean squared error of the parameter estimates, and in identifying the correct dimensionality of encoding using simulated data. We also show that our method provides more accurate inference of low-dimensional subspaces of activity than a competing algorithm, demixed PCA.

BourGAN: Generative Networks with Metric Embeddings

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.

Online Robust Policy Learning in the Presence of Unknown Adversaries

Aaron Havens, Zhanhong Jiang, Soumik Sarkar

The growing prospect of deep reinforcement learning (DRL) being used in cyber-physical systems has raised concerns around safety and robustness of autonomous agents. Recent work on generating adversarial attacks have shown that it is computationally feasible for a bad actor to fool a DRL policy into behaving sub optimally. Although certain adversarial attacks with specific attack models have been addressed, most studies are only interested in off-line optimization in the data space (e.g., example fitting, distillation). This paper introduces a Meta-Learned Advantage Hierarchy (MLAH) framework that is attack model-agnostic and more suited to reinforcement learning, via handling the attacks in the decision space (as opposed to data space) and directly mitigating learned bias introduced by the adversary. In MLAH, we learn separate sub-policies (nominal and adversarial) in an online manner, as guided by a supervisory master agent that detects the presence of the adversary by leveraging the advantage function for the sub-policies. We demonstrate that the proposed algorithm enables policy learning with significantly lower bias as compared to the state-of-the-art policy learning approaches even in the presence of heavy state information attacks. We present algorithm analysis and simulation results using popular OpenAI Gym environments.

Representer Point Selection for Explaining Deep Neural Networks

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, Pradeep K. 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.

Watch Your Step: Learning Node Embeddings via Graph Attention

Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, Alexander A. Alemi

Graph embedding methods represent nodes in a continuous vector space, preserving different types of relational information from the graph.

There are many hyper-parameters to these methods (e.g. the length of a random walk) which have to be manually tuned for every graph.

In this paper, we replace previously fixed hyper-parameters with trainable ones that we automatically learn via backpropagation.

In particular, we propose a novel attention model on the power series of the transition matrix, which guides the random walk to optimize an upstream objective. Unlike previous approaches to attention models, the method that we propose utilizes attention parameters exclusively on the data itself (e.g. on the random walk), and are not used by the model for inference.

We experiment on link prediction tasks, as we aim to produce embeddings that best preserve the graph structure, generalizing to unseen information.

We improve state-of-the-art results on a comprehensive suite of real-world graph datasets including social, collaboration, and biological networks, where we observe that our graph attention model can reduce the error by up to 20%-40%.

We show that our automatically-learned attention parameters can vary significantly per graph, and correspond to the optimal choice of hyper-parameter if we manually tune existing methods.

ℓ_1 -regression with Heavy-tailed Distributions

Lijun Zhang, Zhi-Hua Zhou

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning Confidence Sets using Support Vector Machines

Wenbo Wang, Xingye Qiao

The goal of confidence-set learning in the binary classification setting is to construct two sets, each with a specific probability guarantee to cover a class. An observation outside the overlap of the two sets is deemed to be from one of the two classes, while the overlap is an ambiguity region which could belong to either class. Instead of plug-in approaches, we propose a support vector classifier to construct confidence sets in a flexible manner. Theoretically, we show that the proposed learner can control the non-coverage rates and minimize the ambiguity with high probability. Efficient algorithms are developed and numerical studies illustrate the effectiveness of the proposed method.

Learning to Optimize Tensor Programs

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.

Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation

n

JING LI, Rafal Mantiuk, Junle Wang, Suiyi Ling, Patrick Le Callet

In this paper we present a hybrid active sampling strategy for pairwise preference aggregation, which aims at recovering the underlying rating of the test candidates from sparse and noisy pairwise labeling. Our method employs Bayesian optimization framework and Bradley-Terry model to construct the utility function, then to obtain the Expected Information Gain (EIG) of each pair. For computational efficiency, Gaussian-Hermite quadrature is used for estimation of EIG. In this work, a hybrid active sampling strategy is proposed, either using Global Maximum (GM) EIG sampling or Minimum Spanning Tree (MST) sampling in each trial, which is determined by the test budget. The proposed method has been validated on both simulated and real-world datasets, where it shows higher preference aggregation ability than the state-of-the-art methods.

A no-regret generalization of hierarchical softmax to extreme multi-label classification

Marek Wydmuch, Kalina Jasinska, Mikhail Kuznetsov, Róbert Busa-Fekete, Krzysztof Dembczynski

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

The Sample Complexity of Semi-Supervised Learning with Nonparametric Mixture Models

Chen Dan, Liu Leqi, Bryon Aragam, Pradeep K. Ravikumar, Eric P. Xing

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Estimating Learnability in the Sublinear Data Regime

Weihaio Kong, Gregory Valiant

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Multi-armed Bandits with Compensation

Siwei Wang, Longbo Huang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

A Neural Compositional Paradigm for Image Captioning

Bo Dai, Sanja Fidler, Dahua Lin

Mainstream captioning models often follow a sequential structure to generate cap-

tions, leading to issues such as introduction of irrelevant semantics, lack of diversity in the generated captions, and inadequate generalization performance. In this paper, we present an alternative paradigm for image captioning, which factorizes the captioning procedure into two stages: (1) extracting an explicit semantic representation from the given image; and (2) constructing the caption based on a recursive compositional procedure in a bottom-up manner. Compared to conventional ones, our paradigm better preserves the semantic content through an explicit factorization of semantics and syntax. By using the compositional generation procedure, caption construction follows a recursive structure, which naturally fits the properties of human language. Moreover, the proposed compositional procedure requires less data to train, generalizes better, and yields more diverse

captions.

Stochastic Composite Mirror Descent: Optimal Bounds with High Probabilities

Yunwen Lei, Ke Tang

We study stochastic composite mirror descent, a class of scalable algorithms able to exploit the geometry and composite structure of a problem. We consider both convex and strongly convex objectives with non-smooth loss functions, for each of which we establish high-probability convergence rates optimal up to a logarithmic factor. We apply the derived computational error bounds to study the generalization performance of multi-pass stochastic gradient descent (SGD) in a non-parametric setting. Our high-probability generalization bounds enjoy a logarithmic dependency on the number of passes provided that the step size sequence is square-summable, which improves the existing bounds in expectation with a polynomial dependency and therefore gives a strong justification on the ability of multi-pass SGD to overcome overfitting. Our analysis removes boundedness assumptions on subgradients often imposed in the literature. Numerical results are reported to support our theoretical findings.

Geometry-Aware Recurrent Neural Networks for Active Visual Recognition

Ricson Cheng, Ziyang Wang, Katerina Fragkiadaki

We present recurrent geometry-aware neural networks that integrate visual information 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.

Reward learning from human preferences and demonstrations in Atari

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, Dario Amodei

To solve complex real-world problems with reinforcement learning, we cannot rely on manually specified reward functions. Instead, we need humans to communicate an objective to the agent directly. In this work, we combine two approaches to this problem: learning from expert demonstrations and learning from trajectory preferences. We use both to train a deep neural network to model the reward function and use its predicted reward to train a DQN-based deep reinforcement learning agent on 9 Atari games. Our approach beats the imitation learning baseline in 7 games and achieves strictly superhuman performance on 2 games. Additionally, we investigate the fit of the reward model, present some reward hacking problems, and study the effects of noise in the human labels.

Rectangular Bounding Process

Xuhui Fan, Bin Li, Scott Sisson

Stochastic partition models divide a multi-dimensional space into a number of rectangular regions, such that the data within each region exhibit certain types of homogeneity. Due to the nature of their partition strategy, existing partition models may create many unnecessary divisions in sparse regions when trying to describe data in dense regions. To avoid this problem we introduce a new parsimonious partition model -- the Rectangular Bounding Process (RBP) -- to efficiently partition multi-dimensional spaces, by employing a bounding strategy to enclose

data points within rectangular bounding boxes. Unlike existing approaches, the RBP possesses several attractive theoretical properties that make it a powerful nonparametric partition prior on a hypercube. In particular, the RBP is self-consistent and as such can be directly extended from a finite hypercube to infinite (unbounded) space. We apply the RBP to regression trees and relational models as a flexible partition prior. The experimental results validate the merit of the RBP {in rich yet parsimonious expressiveness} compared to the state-of-the-art methods.

Constructing Unrestricted Adversarial Examples with Generative Models

Yang Song, Rui Shu, Nate Kushman, Stefano Ermon

Adversarial examples are typically constructed by perturbing an existing data point within a small matrix norm, and current defense methods are focused on guarding against this type of attack. In this paper, we propose a new class of adversarial examples that are synthesized entirely from scratch using a conditional generative model, without being restricted to norm-bounded perturbations. We first train an Auxiliary Classifier Generative Adversarial Network (AC-GAN) to model the class-conditional distribution over data samples. Then, conditioned on a desired class, we search over the AC-GAN latent space to find images that are likely under the generative model and are misclassified by a target classifier. We demonstrate through human evaluation that these new kind of adversarial images, which we call Generative Adversarial Examples, are legitimate and belong to the desired class. Our empirical results on the MNIST, SVHN, and CelebA datasets show that generative adversarial examples can bypass strong adversarial training and certified defense methods designed for traditional adversarial attacks.

Causal Discovery from Discrete Data using Hidden Compact Representation

Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, Zhifeng Hao

Causal discovery from a set of observations is one of the fundamental problems across several disciplines. For continuous variables, recently a number of causal discovery methods have demonstrated their effectiveness in distinguishing the cause from effect by exploring certain properties of the conditional distribution, but causal discovery on categorical data still remains to be a challenging problem, because it is generally not easy to find a compact description of the causal mechanism for the true causal direction. In this paper we make an attempt to find a way to solve this problem by assuming a two-stage causal process: the first stage maps the cause to a hidden variable of a lower cardinality, and the second stage generates the effect from the hidden representation. In this way, the causal mechanism admits a simple yet compact representation. We show that under this model, the causal direction is identifiable under some weak conditions on the true causal mechanism. We also provide an effective solution to recover the above hidden compact representation within the likelihood framework. Empirical studies verify the effectiveness of the proposed approach on both synthetic and real-world data.

Boosted Sparse and Low-Rank Tensor Regression

Lifang He, Kun Chen, Wanwan Xu, Jiayu Zhou, Fei Wang

We propose a sparse and low-rank tensor regression model to relate a univariate outcome to a feature tensor, in which each unit-rank tensor from the CP decomposition of the coefficient tensor is assumed to be sparse. This structure is both parsimonious and highly interpretable, as it implies that the outcome is related to the features through a few distinct pathways, each of which may only involve subsets of feature dimensions. We take a divide-and-conquer strategy to simplify the task into a set of sparse unit-rank tensor regression problems. To make the computation efficient and scalable, for the unit-rank tensor regression, we propose a stagewise estimation procedure to efficiently trace out its entire solution path. We show that as the step size goes to zero, the stagewise solution paths converge exactly to those of the corresponding regularized regression. The superior performance of our approach is demonstrated on various real-world and synthetic examples.

Combinatorial Optimization with Graph Convolutional Networks and Guided Tree Search

Zhuwen Li, Qifeng Chen, Vladlen Koltun

We present a learning-based approach to computing solutions for certain NP-hard problems. Our approach combines deep learning techniques with useful algorithmic elements from classic heuristics. The central component is a graph convolutional network that is trained to estimate the likelihood, for each vertex in a graph, of whether this vertex is part of the optimal solution. The network is designed and trained to synthesize a diverse set of solutions, which enables rapid exploration of the solution space via tree search. The presented approach is evaluated on four canonical NP-hard problems and five datasets, which include benchmark satisfiability problems and real social network graphs with up to a hundred thousand nodes. Experimental results demonstrate that the presented approach substantially outperforms recent deep learning work, and performs on par with highly optimized state-of-the-art heuristic solvers for some NP-hard problems. Experiments indicate that our approach generalizes across datasets, and scales to graphs that are orders of magnitude larger than those used during training.

Chaining Mutual Information and Tightening Generalization Bounds

Amir Asadi, Emmanuel Abbe, Sergio Verdu

Bounding the generalization error of learning algorithms has a long history, which yet falls short in explaining various generalization successes including those of deep learning. Two important difficulties are (i) exploiting the dependencies between the hypotheses, (ii) exploiting the dependence between the algorithm's input and output. Progress on the first point was made with the chaining method, originating from the work of Kolmogorov, and used in the VC-dimension bound. More recently, progress on the second point was made with the mutual information method by Russo and Zou '15. Yet, these two methods are currently disjoint. In this paper, we introduce a technique to combine chaining and mutual information methods, to obtain a generalization bound that is both algorithm-dependent and that exploits the dependencies between the hypotheses. We provide an example in which our bound significantly outperforms both the chaining and the mutual information bounds. As a corollary, we tighten Dudley's inequality when the learning algorithm chooses its output from a small subset of hypotheses with high probability.

Modeling Dynamic Missingness of Implicit Feedback for Recommendation

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 *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 *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 an essential underlying factor to subsequent missingness. To model and exploit the dynamics of missingness, we propose a latent variable named *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 incorporates it into matrix factorization (MF) for recommendation. We also explore two types of constraints to achieve a more compact and interpretable representation of *user intents*. Experiments on real-world datasets demonstrate the superiority of our method against state-of-the-art recommender systems.

Does mitigating ML's impact disparity require treatment disparity?

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.

The Sparse Manifold Transform

Yubei Chen, Dylan Paiton, Bruno Olshausen

We present a signal representation framework called the sparse manifold transform that combines key ideas from sparse coding, manifold learning, and slow feature analysis. It turns non-linear transformations in the primary sensory signal space into linear interpolations in a representational embedding space while maintaining approximate invertibility. The sparse manifold transform is an unsupervised and generative framework that explicitly and simultaneously models the sparse discreteness and low-dimensional manifold structure found in natural scenes. When stacked, it also models hierarchical composition. We provide a theoretical description of the transform and demonstrate properties of the learned representation on both synthetic data and natural videos.

Probabilistic Model-Agnostic Meta-Learning

Chelsea Finn, Kelvin Xu, Sergey Levine

Meta-learning for few-shot learning entails acquiring a prior over previous tasks and experiences, such that new tasks be learned from small amounts of data. However, a critical challenge in few-shot learning is task ambiguity: even when a powerful prior can be meta-learned from a large number of prior tasks, a small dataset for a new task can simply be too ambiguous to acquire a single model (e.g., a classifier) for that task that is accurate. In this paper, we propose a probabilistic meta-learning algorithm that can sample models for a new task from a model distribution. Our approach extends model-agnostic meta-learning, which adapts to new tasks via gradient descent, to incorporate a parameter distribution that is trained via a variational lower bound. At meta-test time, our algorithm adapts via a simple procedure that injects noise into gradient descent, and at meta-training time, the model is trained such that this stochastic adaptation procedure produces samples from the approximate model posterior. Our experimental results show that our method can sample plausible classifiers and regressors in ambiguous few-shot learning problems. We also show how reasoning about ambiguity can also be used for downstream active learning problems.

Deep Neural Networks with Box Convolutions

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.

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.

Learning Compressed Transforms with Low Displacement Rank

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.

Deep Defense: Training DNNs with Improved Adversarial Robustness

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 data sets (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>.

Precision and Recall for Time Series

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.

Neighbourhood Consensus Networks

Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, Josef Sivic

We address the problem of finding reliable dense correspondences between a pair of images. This is a challenging task due to strong appearance differences between the corresponding scene elements and ambiguities generated by repetitive patterns. The contributions of this work are threefold. First, inspired by the classic idea of disambiguating feature matches using semi-local constraints, we develop an end-to-end trainable convolutional neural network architecture that identifies sets of spatially consistent matches by analyzing neighbourhood consensus patterns in the 4D space of all possible correspondences between a pair of images.

es without the need for a global geometric model. Second, we demonstrate that the model can be trained effectively from weak supervision in the form of matching and non-matching image pairs without the need for costly manual annotation of point to point correspondences.

Third, we show the proposed neighbourhood consensus network can be applied to a range of matching tasks including both category- and instance-level matching, obtaining the state-of-the-art results on the PF Pascal dataset and the InLoc indoor visual localization benchmark.

PAC-learning in the presence of adversaries

Daniel Cullina, Arjun Nitin Bhagoji, Prateek Mittal

The existence of evasion attacks during the test phase of machine learning algorithms represents a significant challenge to both their deployment and understanding. These attacks can be carried out by adding imperceptible perturbations to inputs to generate adversarial examples and finding effective defenses and detectors has proven to be difficult. In this paper, we step away from the attack-defense arms race and seek to understand the limits of what can be learned in the presence of an evasion adversary. In particular, we extend the Probably Approximately Correct (PAC)-learning framework to account for the presence of an adversary. We first define corrupted hypothesis classes which arise from standard binary hypothesis classes in the presence of an evasion adversary and derive the Vapnik-Chervonenkis (VC)-dimension for these, denoted as the adversarial VC-dimension.

We then show that sample complexity upper bounds from the Fundamental Theorem of Statistical learning can be extended to the case of evasion adversaries, where the sample complexity is controlled by the adversarial VC-dimension. We then explicitly derive the adversarial VC-dimension for halfspace classifiers in the presence of a sample-wise norm-constrained adversary of the type commonly studied for evasion attacks and show that it is the same as the standard VC-dimension, closing an open question. Finally, we prove that the adversarial VC-dimension can be either larger or smaller than the standard VC-dimension depending on the hypothesis class and adversary, making it an interesting object of study in its own right.

Optimal Algorithms for Non-Smooth Distributed Optimization in Networks

Kevin Scaman, Francis Bach, Sebastien Bubeck, Laurent Massoulié, Yin Tat Lee

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Large-Scale Stochastic Sampling from the Probability Simplex

Jack Baker, Paul Fearnhead, Emily Fox, Christopher Nemeth

Stochastic gradient Markov chain Monte Carlo (SGMCMC) has become a popular method for scalable Bayesian inference. These methods are based on sampling a discrete-time approximation to a continuous time process, such as the Langevin diffusion. When applied to distributions defined on a constrained space the time-discretization error can dominate when we are near the boundary of the space. We demonstrate that because of this, current SGMCMC methods for the simplex struggle with sparse simplex spaces; when many of the components are close to zero. Unfortunately, many popular large-scale Bayesian models, such as network or topic models, require inference on sparse simplex spaces. To avoid the biases caused by this discretization error, we propose the stochastic Cox-Ingersoll-Ross process (SCIR), which removes all discretization error and we prove that samples from the SCIR process are asymptotically unbiased. We discuss how this idea can be extended to target other constrained spaces. Use of the SCIR process within a SGMCMC algorithm is shown to give substantially better performance for a topic model and a Dirichlet process mixture model than existing SGMCMC approaches.

Transfer of Value Functions via Variational Methods

Andrea Tirinzoni, Rafael Rodriguez Sanchez, Marcello Restelli

We consider the problem of transferring value functions in reinforcement learning. We propose an approach that uses the given source tasks to learn a prior distribution over optimal value functions and provide an efficient variational approximation of the corresponding posterior in a new target task. We show our approach to be general, in the sense that it can be combined with complex parametric function approximators and distribution models, while providing two practical algorithms based on Gaussians and Gaussian mixtures. We theoretically analyze them by deriving a finite-sample analysis and provide a comprehensive empirical evaluation in four different domains.

Adaptive Methods for Nonconvex Optimization

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, Sanjiv Kumar

Adaptive gradient methods that rely on scaling gradients down by the square root of exponential moving averages of past squared gradients, such as RMSProp, Adam, and Adadelta have found wide application in optimizing the nonconvex problems that arise in deep learning. However, it has been recently demonstrated that such methods can fail to converge even in simple convex optimization settings. In this work, we provide a new analysis of such methods applied to nonconvex stochastic optimization problems, characterizing the effect of increasing minibatch size. Our analysis shows that under this scenario such methods do converge to stationarity up to the statistical limit of variance in the stochastic gradients (scaled by a constant factor). In particular, our result implies that increasing minibatch sizes enables convergence, thus providing a way to circumvent the non-convergence issues. Furthermore, we provide a new adaptive optimization algorithm, Yogi, which controls the increase in effective learning rate, leading to even better performance with similar theoretical guarantees on convergence. Extensive experiments show that Yogi with very little hyperparameter tuning outperforms methods such as Adam in several challenging machine learning tasks.

How Does Batch Normalization Help Optimization?

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.

Compact Generalized Non-local Network

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>.

Stacked Semantics-Guided Attention Model for Fine-Grained Zero-Shot Learning

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.

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

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.

Multiplicative Weights Updates with Constant Step-Size in Graphical Constant-Sum Games

Yun Kuen Cheung

Since Multiplicative Weights (MW) updates are the discrete analogue of the continuous Replicator Dynamics (RD), some researchers had expected their qualitative behaviours would be similar. We show that this is false in the context of graphical constant-sum games, which include two-person zero-sum games as special cases. In such games which have a fully-mixed Nash Equilibrium (NE), it was known that RD satisfy the permanence and Poincare recurrence properties, but we show that MW updates with any constant step-size $\epsilon > 0$ converge to the boundary of the state space, and thus do not satisfy the two properties. Using this result, we show that MW updates have a regret lower bound of $\Omega(1 / (\epsilon T))$, while it was known that the regret of RD is upper bounded by $O(1 / T)$.

Efficient Online Portfolio with Logarithmic Regret

Haipeng Luo, Chen-Yu Wei, Kai Zheng

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Banach Wasserstein GAN

Jonas Adler, Sebastian Lunz

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

SplineNets: Continuous Neural Decision Graphs

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.

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

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.

Implicit Reparameterization Gradients

Mikhail Figurnov, Shakir Mohamed, Andriy Mnih

By providing a simple and efficient way of computing low-variance gradients of continuous random variables, the reparameterization trick has become the technique of choice for training a variety of latent variable models. However, it is not applicable to a number of important continuous distributions. We introduce an alternative approach to computing reparameterization gradients based on implicit differentiation and demonstrate its broader applicability by applying it to Gamma, Beta, Dirichlet, and von Mises distributions, which cannot be used with the classic reparameterization trick. Our experiments show that the proposed approach is faster and more accurate than the existing gradient estimators for these distributions.

Visual Object Networks: Image Generation with Disentangled 3D Representations

Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh T

enenbaum, Bill Freeman

Recent progress in deep generative models has led to tremendous breakthroughs in image generation. While being able to synthesize photorealistic images, existing models lack an understanding of our underlying 3D world. Different from previous works built on 2D datasets and models, we present a new generative model, Visual Object Networks (VONs), synthesizing natural images of objects with a disentangled 3D representation. Inspired by classic graphics rendering pipelines, we unravel the image formation process into three conditionally independent factors--shape, viewpoint, and texture---and present an end-to-end adversarial learning framework that jointly models 3D shape and 2D texture. Our model first learns to synthesize 3D shapes that are indistinguishable from real shapes. It then renders the object's 2.5D sketches (i.e., silhouette and depth map) from its shape under a sampled viewpoint. Finally, it learns to add realistic textures to these 2.5D sketches to generate realistic images. The VON not only generates images that are more realistic than the state-of-the-art 2D image synthesis methods but also enables many 3D operations such as changing the viewpoint of a generated image, shape and texture editing, linear interpolation in texture and shape space, and transferring appearance across different objects and viewpoints.

Neural Architecture Optimization

Renqian Luo, Fei Tian, Tao Qin, Enhong Chen, Tie-Yan Liu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare

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.

Learning Optimal Reserve Price against Non-myopic Bidders

Jinyan Liu, Zhiyi Huang, Xiangning Wang

We consider the problem of learning optimal reserve price in repeated auctions against non-myopic bidders, who may bid strategically in order to gain in future rounds even if the single-round auctions are truthful. Previous algorithms, e.g., empirical pricing, do not provide non-trivial regret bounds in this setting in general. We introduce algorithms that obtain small regret against non-myopic bidders either when the market is large, i.e., no bidder appears in a constant fraction of the rounds, or when the bidders are impatient, i.e., they discount future

re utility by some factor mildly bounded away from one. Our approach carefully controls what information is revealed to each bidder, and builds on techniques from differentially private online learning as well as the recent line of works on jointly differentially private algorithms.

A Bayesian Approach to Generative Adversarial Imitation Learning

Wonseok Jeon, Seokin Seo, Kee-Eung Kim

Generative adversarial training for imitation learning has shown promising results on high-dimensional and continuous control tasks. This paradigm is based on reducing the imitation learning problem to the density matching problem, where the agent iteratively refines the policy to match the empirical state-action visitation frequency of the expert demonstration. Although this approach has shown to robustly learn to imitate even with scarce demonstration, one must still address the inherent challenge that collecting trajectory samples in each iteration is a costly operation. To address this issue, we first propose a Bayesian formulation of generative adversarial imitation learning (GAIL), where the imitation policy and the cost function are represented as stochastic neural networks. Then, we show that we can significantly enhance the sample efficiency of GAIL leveraging the predictive density of the cost, on an extensive set of imitation learning tasks with high-dimensional states and actions.

Credit Assignment For Collective Multiagent RL With Global Rewards

Duc Thien Nguyen, Akshat Kumar, Hoong Chuin Lau

Scaling decision theoretic planning to large multiagent systems is challenging due to uncertainty and partial observability in the environment. We focus on a multiagent planning model subclass, relevant to urban settings, where agent interactions are dependent on their "collective influence" on each other, rather than their identities. Unlike previous work, we address a general setting where system reward is not decomposable among agents. We develop collective actor-critic RL approaches for this setting, and address the problem of multiagent credit assignment, and computing low variance policy gradient estimates that result in faster convergence to high quality solutions. We also develop difference rewards based credit assignment methods for the collective setting. Empirically our new approaches provide significantly better solutions than previous methods in the presence of global rewards on two real world problems modeling taxi fleet optimization and multiagent patrolling, and a synthetic grid navigation domain.

Knowledge Distillation by On-the-Fly Native Ensemble

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 favorable 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.

Flexible and accurate inference and learning for deep generative models

Eszter V3rtes, Maneesh Sahani

We introduce a new approach to learning in hierarchical latent-variable generative

models called the "distributed distributional code Helmholtz machine", which emphasises flexibility and accuracy in the inferential process. Like the original

Helmholtz machine and later variational autoencoder algorithms (but unlike adversarial methods) our approach learns an explicit inference or “recognition” model to approximate the posterior distribution over the latent variables. Unlike these earlier methods, it employs a posterior representation that is not limited to a narrow tractable parametrised form (nor is it represented by samples). To train the generative and recognition models we develop an extended wake-sleep algorithm inspired by the original Helmholtz machine. This makes it possible to learn hierarchical latent models with both discrete and continuous variables, where an accurate posterior representation is essential. We demonstrate that the new algorithm outperforms current state-of-the-art methods on synthetic, natural image patch and the MNIST data sets.

A loss framework for calibrated anomaly detection

Given samples from a probability distribution, anomaly detection is the problem of determining if a given point lies in a low-density region. This paper concerns calibrated anomaly detection, which is the practically relevant extension where we additionally wish to produce a confidence score for a point being anomalous. Building on a classification framework for anomaly detection, we show how minimisation of a suitably modified proper loss produces density estimates only for anomalous instances. We then show how to incorporate quantile control by relating our objective to a generalised version of the pinball loss. Finally, we show how to efficiently optimise the objective with kernelised scorer, by leveraging a recent result from the point process literature. The resulting objective captures a close relative of the one-class SVM as a special case.

Persistence Fisher Kernel: A Riemannian Manifold Kernel for Persistence Diagrams
Tam Le, Makoto Yamada

Algebraic topology methods have recently played an important role for statistical analysis with complicated geometric structured data such as shapes, linked twist maps, and material data. Among them, `\textit{persistent homology}` is a well-known tool to extract robust topological features, and outputs as `\textit{persistence diagrams}` (PDs). However, PDs are point multi-sets which can not be used in machine learning algorithms for vector data. To deal with it, an emerged approach is to use kernel methods, and an appropriate geometry for PDs is an important factor to measure the similarity of PDs. A popular geometry for PDs is the `\textit{Wasserstein metric}`. However, Wasserstein distance is not `\textit{negative definite}`. Thus, it is limited to build positive definite kernels upon the Wasserstein distance `\textit{without approximation}`. In this work, we rely upon the alternative `\textit{Fisher information geometry}` to propose a positive definite kernel for PDs `\textit{without approximation}`, namely the Persistence Fisher (PF) kernel. Then, we analyze eigensystem of the integral operator induced by the proposed kernel for kernel machines. Based on that, we derive generalization error bounds via covering numbers and Rademacher averages for kernel machines with the PF kernel. Additionally, we show some nice properties such as stability and infinite divisibility for the proposed kernel. Furthermore, we also propose a linear time complexity over the number of points in PDs for an approximation of our proposed kernel with a bounded error. Throughout experiments with many different tasks on various benchmark datasets, we illustrate that the PF kernel compares favorably with other baseline kernels for PDs.

Constructing Deep Neural Networks by Bayesian Network Structure Learning
Raanan Y. Rohekar, Shami Nisimov, Yaniv Gurwicz, Guy Koren, Gal Novik

We introduce a principled approach for unsupervised structure learning of deep n

neural networks. We propose a new interpretation for depth and inter-layer connectivity where conditional independencies in the input distribution are encoded hierarchically in the network structure. Thus, the depth of the network is determined inherently. The proposed method casts the problem of neural network structure learning as a problem of Bayesian network structure learning. Then, instead of directly learning the discriminative structure, it learns a generative graph, constructs its stochastic inverse, and then constructs a discriminative graph. We prove that conditional-dependency relations among the latent variables in the generative graph are preserved in the class-conditional discriminative graph. We demonstrate on image classification benchmarks that the deepest layers (convolutional and dense) of common networks can be replaced by significantly smaller learned structures, while maintaining classification accuracy---state-of-the-art on tested benchmarks. Our structure learning algorithm requires a small computational cost and runs efficiently on a standard desktop CPU.

Training Deep Models Faster with Robust, Approximate Importance Sampling

Tyler B. 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 descent. 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.

Learning Beam Search Policies via Imitation Learning

Renato Negrinho, Matthew Gormley, Geoffrey J. Gordon

Beam search is widely used for approximate decoding in structured prediction problems. Models often use a beam at test time but ignore its existence at train time, and therefore do not explicitly learn how to use the beam. We develop an unifying meta-algorithm for learning beam search policies using imitation learning.

In our setting, the beam is part of the model and not just an artifact of approximate decoding. Our meta-algorithm captures existing learning algorithms and suggests new ones. It also lets us show novel no-regret guarantees for learning beam search policies.

Multivariate Time Series Imputation with Generative Adversarial Networks

Yonghong Luo, Xiangrui Cai, Ying ZHANG, Jun Xu, Yuan xiaojie

Multivariate time series usually contain a large number of missing values, which hinders the application of advanced analysis methods on multivariate time series data. Conventional approaches to addressing the challenge of missing values, including mean/zero imputation, case deletion, and matrix factorization-based imputation, are all incapable of modeling the temporal dependencies and the nature of complex distribution in multivariate time series. In this paper, we treat the problem of missing value imputation as data generation. Inspired by the success of Generative Adversarial Networks (GAN) in image generation, we propose to learn the overall distribution of a multivariate time series dataset with GAN, which is further used to generate the missing values for each sample. Different from the image data, the time series data are usually incomplete due to the nature of data recording process. A modified Gate Recurrent Unit is employed in GAN to model the temporal irregularity of the incomplete time series. Experiments on two multivariate time series datasets show that the proposed model outperformed the baselines in terms of accuracy of imputation. Experimental results also showed that a simple model on the imputed data can achieve state-of-the-art results on the prediction tasks, demonstrating the benefits of our model in downstream applications.

An Efficient Pruning Algorithm for Robust Isotonic Regression

Cong Han Lim

We study a generalization of the classic isotonic regression problem where we allow separable nonconvex objective functions, focusing on the case of estimators used in robust regression. A simple dynamic programming approach allows us to solve this problem to within ϵ -accuracy (of the global minimum) in time linear in $1/\epsilon$ and the dimension. We can combine techniques from the convex case with branch-and-bound ideas to form a new algorithm for this problem that naturally exploits the shape of the objective function. Our algorithm achieves the best bounds for both the general nonconvex and convex case (linear in $\log(1/\epsilon)$), while performing much faster in practice than a straightforward dynamic programming approach, especially as the desired accuracy increases.

Bilinear Attention Networks

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.

Unsupervised Video Object Segmentation for Deep Reinforcement Learning

Vikash Goel, Jameson Weng, Pascal Poupart

We present a new technique for deep reinforcement learning that automatically detects moving objects and uses the relevant information for action selection. The detection of moving objects is done in an unsupervised way by exploiting structure from motion. Instead of directly learning a policy from raw images, the agent first learns to detect and segment moving objects by exploiting flow information in video sequences. The learned representation is then used to focus the policy of the agent on the moving objects. Over time, the agent identifies which objects are critical for decision making and gradually builds a policy based on relevant moving objects. This approach, which we call Motion-Oriented REinforcement Learning (MOREL), is demonstrated on a suite of Atari games where the ability to detect moving objects reduces the amount of interaction needed with the environment to obtain a good policy. Furthermore, the resulting policy is more interpretable than policies that directly map images to actions or values with a black box neural network. We can gain insight into the policy by inspecting the segmentation and motion of each object detected by the agent. This allows practitioners to confirm whether a policy is making decisions based on sensible information. Our code is available at <https://github.com/vik-goel/MOREL>.

Constructing Fast Network through Deconstruction of Convolution

Yunho Jeon, Junmo Kim

Convolutional neural networks have achieved great success in various vision tasks; however, they incur heavy resource costs. By using deeper and wider networks, network accuracy can be improved rapidly. However, in an environment with limited resources (e.g., mobile applications), heavy networks may not be usable. This study shows that naive convolution can be deconstructed into a shift operation and pointwise convolution. To cope with various convolutions, we propose a new shift operation called active shift layer (ASL) that formulates the amount of shift as a learnable function with shift parameters. This new layer can be optimized end-to-end through backpropagation and it can provide optimal shift values. Fi

nally, we apply this layer to a light and fast network that surpasses existing state-of-the-art networks.

Improving Simple Models with Confidence Profiles

Amit Dhurandhar, Karthikeyan Shanmugam, Ronny Luss, Peder A. Olsen

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Turbo Learning for CaptionBot and DrawingBot

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 pseudo-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.

Online Reciprocal Recommendation with Theoretical Performance Guarantees

Fabio Vitale, Nikos Parotsidis, Claudio Gentile

A reciprocal recommendation problem is one where the goal of learning is not just to predict a user's preference towards a passive item (e.g., a book), but to recommend the targeted user on one side another user from the other side such that a mutual interest between the two exists. The problem thus is sharply different from the more traditional items-to-users recommendation, since a good match requires meeting the preferences of both users. We initiate a rigorous theoretical investigation of the reciprocal recommendation task in a specific framework of sequential learning. We point out general limitations, formulate reasonable assumptions enabling effective learning and, under these assumptions, we design and analyze a computationally efficient algorithm that uncovers mutual likes at a pace comparable to those achieved by a clairvoyant algorithm knowing all user preferences in advance. Finally, we validate our algorithm against synthetic and real-world datasets, showing improved empirical performance over simple baselines.

Learning semantic similarity in a continuous space

Michel Deudon

We address the problem of learning semantic representation of questions to measure similarity between pairs as a continuous distance metric. Our work naturally extends Word Mover's Distance (WMD) [1] by representing text documents as normal distributions instead of bags of embedded words. Our learned metric measures the dissimilarity between two questions as the minimum amount of distance the intent (hidden representation) of one question needs to "travel" to match the intent of another question. We first learn to repeat, reformulate questions to infer intents as normal distributions with a deep generative model [2] (variational auto encoder). Semantic similarity between pairs is then learned discriminatively as an optimal transport distance metric (Wasserstein 2) with our novel variational siamese framework. Among known models that can read sentences individually, our

Our proposed framework achieves competitive results on Quora duplicate questions dataset. Our work sheds light on how deep generative models can approximate distributions (semantic representations) to effectively measure semantic similarity with meaningful distance metrics from Information Theory.

Representation Balancing MDPs for Off-policy Policy Evaluation

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A. Faisal, Finale Doshi-Velez, Emma Brunskill

We study the problem of off-policy policy evaluation (OPPE) in RL. In contrast to prior work, we consider how to estimate both the individual policy value and an average policy value accurately. We draw inspiration from recent work in causal reasoning, and propose a new finite sample generalization error bound for value estimates from MDP models. Using this upper bound as an objective, we develop a learning algorithm of an MDP model with a balanced representation, and show that our approach can yield substantially lower MSE in common synthetic benchmarks and a HIV treatment simulation domain.

See and Think: Disentangling Semantic Scene Completion

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 inputting 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 data sets.

L4: Practical loss-based stepsize adaptation for deep learning

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.

Generalisation of structural knowledge in the hippocampal-entorhinal system

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.

nctions 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.

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

Robert J. Wang, Xiang Li, Charles X. 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.

Co-regularized Alignment for Unsupervised Domain Adaptation

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.

How To Make the Gradients Small Stochastically: Even Faster Convex and Nonconvex SGD

Zeyuan Allen-Zhu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Entropy Rate Estimation for Markov Chains with Large State Space

Yanjun Han, Jiantao Jiao, Chuan-Zheng Lee, Tsachy Weissman, Yihong Wu, Tiancheng Yu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Data-dependent PAC-Bayes priors via differential privacy

Gintare Karolina Dziugaite, Daniel M. Roy

The Probably Approximately Correct (PAC) Bayes framework (McAllester, 1999) can incorporate knowledge about the learning algorithm and (data) distribution through the use of distribution-dependent priors, yielding tighter generalization bounds on data-dependent posteriors. Using this flexibility, however, is difficult, especially when the data distribution is presumed to be unknown. We show how a differentially private data-dependent prior yields a valid PAC-Bayes bound, and then show how non-private mechanisms for choosing priors can also yield generalization bounds. As an application of this result, we show that a Gaussian prior mean chosen via stochastic gradient Langevin dynamics (SGLD; Welling and Teh, 2011) leads to a valid PAC-Bayes bound due to control of the 2-Wasserstein distance to a differentially private stationary distribution. We study our data-dependent bounds empirically, and show that they can be nonvacuous even when other distribution-dependent bounds are vacuous.

Unsupervised Depth Estimation, 3D Face Rotation and Replacement

Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, Chris Pal

We present an unsupervised approach for learning to estimate three dimensional (3D) facial structure from a single image while also predicting 3D viewpoint transformations that match a desired pose and facial geometry.

We achieve this by inferring the depth of facial keypoints of an input image in an unsupervised manner, without using any form of ground-truth depth information. We show how it is possible to use these depths as intermediate computations within a new backpropable loss to predict the parameters of a 3D affine transformation matrix that maps inferred 3D keypoints of an input face to the corresponding 2D keypoints on a desired target facial geometry or pose.

Our resulting approach, called DepthNets, can therefore be used to infer plausible 3D transformations from one face pose to another, allowing faces to be frontalized, transformed into 3D models or even warped to another pose and facial geometry.

Lastly, we identify certain shortcomings with our formulation, and explore adversarial image translation techniques as a post-processing step to re-synthesize complete head shots for faces re-targeted to different poses or identities.

Information Constraints on Auto-Encoding Variational Bayes

Romain Lopez, Jeffrey Regier, Michael I. Jordan, Nir Yosef

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

On Misinformation Containment in Online Social Networks

Amo Tong, Ding-Zhu Du, Weili Wu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Autoconj: Recognizing and Exploiting Conjugacy Without a Domain-Specific Language

Matthew D. Hoffman, Matthew J. Johnson, Dustin Tran

Deriving conditional and marginal distributions using conjugacy relationships can be time consuming and error prone. In this paper, we propose a strategy for automating such derivations. Unlike previous systems which focus on relationships between pairs of random variables, our system (which we call Autoconj) operates directly on Python functions that compute log-joint distribution functions. Autoconj provides support for conjugacy-exploiting algorithms in any Python-embedded PPL. This paves the way for accelerating development of novel inference algorithms and structure-exploiting modeling strategies. The package can be downloaded at <https://github.com/google-research/autoconj>.

Size-Noise Tradeoffs in Generative Networks

Bolton Bailey, Matus J. Telgarsky

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization

Pan Xu, Jinghui Chen, Difan Zou, Quanquan Gu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Online convex optimization for cumulative constraints

Jianjun Yuan, Andrew Lamperski

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Semi-supervised Deep Kernel Learning: Regression with Unlabeled Data by Minimizing Predictive Variance

Neal Jean, Sang Michael Xie, Stefano Ermon

Large amounts of labeled data are typically required to train deep learning models. For many real-world problems, however, acquiring additional data can be expensive or even impossible. We present semi-supervised deep kernel learning (SSDKL), a semi-supervised regression model based on minimizing predictive variance in the posterior regularization framework. SSDKL combines the hierarchical representation learning of neural networks with the probabilistic modeling capabilities of Gaussian processes. By leveraging unlabeled data, we show improvements on a diverse set of real-world regression tasks over supervised deep kernel learning and semi-supervised methods such as VAT and mean teacher adapted for regression.

.

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

Fabian Sinz, Alexander S. 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 pro

vide 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.

Sigsoftmax: Reanalysis of the Softmax Bottleneck

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.

Parsimonious Quantile Regression of Financial Asset Tail Dynamics via Sequential Learning

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.

Adaptive Learning with Unknown Information Flows

Yonatan Gur, Ahmadreza Momeni

An agent facing sequential decisions that are characterized by partial feedback needs to strike a balance between maximizing immediate payoffs based on available information, and acquiring new information that may be essential for maximizing future payoffs. This trade-off is captured by the multi-armed bandit (MAB) framework that has been studied and applied when at each time epoch payoff observations are collected on the actions that are selected at that epoch. In this paper we introduce a new, generalized MAB formulation in which additional information on each arm may appear arbitrarily throughout the decision horizon, and study the impact of such information flows on the achievable performance and the design of efficient decision-making policies. By obtaining matching lower and upper bounds, we characterize the (regret) complexity of this family of MAB problems as a function of the information flows. We introduce an adaptive exploration policy that, without any prior knowledge of the information arrival process, attains the best performance (in terms of regret rate) that is achievable when the information arrival process is a priori known. Our policy uses dynamically customized

virtual time indexes to endogenously control the exploration rate based on the realized information arrival process.

DVAE#: Discrete Variational Autoencoders with Relaxed Boltzmann Priors

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.

Reinforcement Learning for Solving the Vehicle Routing Problem

MohammadReza Nazari, Afshin Oroojlooy, Lawrence Snyder, Martin Takac

We present an end-to-end framework for solving the Vehicle Routing Problem (VRP) using reinforcement learning. In this approach, we train a single policy model that finds near-optimal solutions for a broad range of problem instances of similar size, only by observing the reward signals and following feasibility rules. We consider a parameterized stochastic policy, and by applying a policy gradient algorithm to optimize its parameters, the trained model produces the solution as a sequence of consecutive actions in real time, without the need to re-train for every new problem instance. On capacitated VRP, our approach outperforms classical heuristics and Google's OR-Tools on medium-sized instances in solution quality with comparable computation time (after training). We demonstrate how our approach can handle problems with split delivery and explore the effect of such deliveries on the solution quality. Our proposed framework can be applied to other variants of the VRP such as the stochastic VRP, and has the potential to be applied more generally to combinatorial optimization problems

GumBolt: Extending Gumbel trick to Boltzmann priors

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 multimodal 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.

Adding One Neuron Can Eliminate All Bad Local Minima

SHIYU LIANG, Ruoyu Sun, Jason D. Lee, R. Srikant

One of the main difficulties in analyzing neural networks is the non-convexity of the loss function which may have many bad local minima. In this paper, we study the landscape of neural networks for binary classification tasks. Under mild assumptions, we prove that after adding one special neuron with a skip connection to the output, or one special neuron per layer, every local minimum is a global minimum.

Norm matters: efficient and accurate normalization schemes in deep networks

Elad Hoffer, Ron Banner, Itay Golan, Daniel Soudry

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Local Differential Privacy for Evolving Data

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.

Dialog-based Interactive Image Retrieval

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.

Byzantine Stochastic Gradient Descent

Dan Alistarh, Zeyuan Allen-Zhu, Jerry Li

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Robust Hypothesis Testing Using Wasserstein Uncertainty Sets

RUI GAO, Liyan Xie, Yao Xie, Huan Xu

We develop a novel computationally efficient and general framework for robust hypothesis testing. The new framework features a new way to construct uncertainty sets under the null and the alternative distributions, which are sets centered around the empirical distribution defined via Wasserstein metric, thus our approach is data-driven and free of distributional assumptions. We develop a convex safe approximation of the minimax formulation and show that such approximation renders a nearly-optimal detector among the family of all possible tests. By exploiting the structure of the least favorable distribution, we also develop a tractable reformulation of such approximation, with complexity independent of the dimension of observation space and can be nearly sample-size-independent in general.

Real-data example using human activity data demonstrated the excellent performance of the new robust detector.

Life-Long Disentangled Representation Learning with Cross-Domain Latent Homology

es

Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, Irina Higgins

Intelligent behaviour in the real-world requires the ability to acquire new knowledge from an ongoing sequence of experiences while preserving and reusing past knowledge. We propose a novel algorithm for unsupervised representation learning from piece-wise stationary visual data: Variational Autoencoder with Shared Embeddings (VASE). Based on the Minimum Description Length principle, VASE automatically detects shifts in the data distribution and allocates spare representational capacity to new knowledge, while simultaneously protecting previously learnt representations from catastrophic forgetting. Our approach encourages the learnt representations to be disentangled, which imparts a number of desirable properties: VASE can deal sensibly with ambiguous inputs, it can enhance its own representations through imagination-based exploration, and most importantly, it exhibits semantically meaningful sharing of latents between different datasets. Compared to baselines with entangled representations, our approach is able to reason beyond surface-level statistics and perform semantically meaningful cross-domain inference.

Computationally and statistically efficient learning of causal Bayes nets using path queries

Kevin Bello, Jean Honorio

Causal discovery from empirical data is a fundamental problem in many scientific domains. Observational data allows for identifiability only up to Markov equivalence class. In this paper we first propose a polynomial time algorithm for learning the exact correctly-oriented structure of the transitive reduction of any causal Bayesian network with high probability, by using interventional path queries. Each path query takes as input an origin node and a target node, and answers whether there is a directed path from the origin to the target. This is done by intervening on the origin node and observing samples from the target node. We theoretically show the logarithmic sample complexity for the size of interventional data per path query, for continuous and discrete networks. We then show how to learn the transitive edges using also logarithmic sample complexity (albeit in time exponential in the maximum number of parents for discrete networks), which allows us to learn the full network. We further extend our work by reducing the number of interventional path queries for learning rooted trees. We also provide an analysis of imperfect interventions.

Predictive Approximate Bayesian Computation via Saddle Points

Yingxiang Yang, Bo Dai, Negar Kiyavash, Niao He

Approximate Bayesian computation (ABC) is an important methodology for Bayesian inference when the likelihood function is intractable. Sampling-based ABC algorithms such as rejection- and K2-ABC are inefficient when the parameters have high dimensions, while the regression-based algorithms such as K- and DR-ABC are hard to scale. In this paper, we introduce an optimization-based ABC framework that addresses these deficiencies. Leveraging a generative model for posterior and joint distribution matching, we show that ABC can be framed as saddle point problems, whose objectives can be accessed directly with samples. We present the predictive ABC algorithm (P-ABC), and provide a probabilistically approximately correct (PAC) bound that guarantees its learning consistency. Numerical experiments show that P-ABC outperforms both K2- and DR-ABC significantly.

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

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

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.

Smoothed analysis of the low-rank approach for smooth semidefinite programs

Thomas Pumić, Samy Jelassi, Nicolas Boumal

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Differentially Private Contextual Linear Bandits

Roshan Shariff, Or Sheffet

We study the contextual linear bandit problem, a version of the standard stochastic multi-armed bandit (MAB) problem where a learner sequentially selects actions to maximize a reward which depends also on a user provided per-round context. Though the context is chosen arbitrarily or adversarially, the reward is assumed to be a stochastic function of a feature vector that encodes the context and selected action. Our goal is to devise private learners for the contextual linear bandit problem.

Learning to Reason with Third Order Tensor Products

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.

Diversity-Driven Exploration Strategy for Deep Reinforcement Learning

Zhang-Wei Hong, Tzu-Yun Shann, Shih-Yang Su, Yi-Hsiang Chang, Tsu-Jui Fu, Chun-Y

i Lee

Efficient exploration remains a challenging research problem in reinforcement learning, especially when an environment contains large state spaces, deceptive local optima, or sparse rewards.

To tackle this problem, we present a diversity-driven approach for exploration, which can be easily combined with both off- and on-policy reinforcement learning algorithms. We show that by simply adding a distance measure to the loss function, the proposed methodology significantly enhances an agent's exploratory behaviors, and thus preventing the policy from being trapped in local optima. We further propose an adaptive scaling method for stabilizing the learning process. We demonstrate the effectiveness of our method in huge 2D gridworlds and a variety of benchmark environments, including Atari 2600 and MuJoCo. Experimental results show that our method outperforms baseline approaches in most tasks in terms of mean scores and exploration efficiency.

On Neuronal Capacity

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.

GroupReduce: Block-Wise Low-Rank Approximation for Neural Language Model Shrinking

Patrick Chen, Si Si, Yang Li, Ciprian Chelba, Cho-Jui Hsieh

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Deep Structured Prediction with Nonlinear Output Transformations

Colin Graber, Ofer Meshi, Alexander Schwing

Deep structured models are widely used for tasks like semantic segmentation, where explicit correlations between variables provide important prior information which generally helps to reduce the data needs of deep nets. However, current deep structured models are restricted by oftentimes very local neighborhood structure, which cannot be increased for computational complexity reasons, and by the fact that the output configuration, or a representation thereof, cannot be transformed further. Very recent approaches which address those issues include graphical model inference inside deep nets so as to permit subsequent non-linear output space transformations. However, optimization of those formulations is challenging and not well understood. Here, we develop a novel model which generalizes existing approaches, such as structured prediction energy networks, and discuss a formulation which maintains applicability of existing inference techniques.

Training Neural Networks Using Features Replay

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

y, 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.

Mallows Models for Top-k Lists

Flavio Chierichetti, Anirban Dasgupta, Shahrzad Haddadan, Ravi Kumar, Silvio Lattanzi

The classic Mallows model is a widely-used tool to realize distributions on permutations. Motivated by common practical situations, in this paper, we generalize Mallows to model distributions on top-k lists by using a suitable distance measure between top-k lists. Unlike many earlier works, our model is both analytically tractable and computationally efficient. We demonstrate this by studying two basic problems in this model, namely, sampling and reconstruction, from both an algorithmic and experimental points of view.

Information-based Adaptive Stimulus Selection to Optimize Communication Efficiency in Brain-Computer Interfaces

Boyla Mainsah, Dmitry Kalika, Leslie Collins, Siyuan Liu, Chandra Throckmorton
Stimulus-driven brain-computer interfaces (BCIs), such as the P300 speller, rely on using a sequence of sensory stimuli to elicit specific neural responses as control signals, while a user attends to relevant target stimuli that occur within the sequence. In current BCIs, the stimulus presentation schedule is typically generated in a pseudo-random fashion. Given the non-stationarity of brain electrical signals, a better strategy could be to adapt the stimulus presentation schedule in real-time by selecting the optimal stimuli that will maximize the signal-to-noise ratios of the elicited neural responses and provide the most information about the user's intent based on the uncertainties of the data being measured. However, the high-dimensional stimulus space limits the development of algorithms with tractable solutions for optimized stimulus selection to allow for real-time decision-making within the stringent time requirements of BCI processing. We derive a simple analytical solution of an information-based objective function for BCI stimulus selection by transforming the high-dimensional stimulus space into a one-dimensional space that parameterizes the objective function - the prior probability mass of the stimulus under consideration, irrespective of its contents. We demonstrate the utility of our adaptive stimulus selection algorithm in improving BCI performance with results from simulation and real-time human experiments.

Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β -Divergences

Jeremias Knoblauch, Jack E. Jewson, Theodoros Damoulas

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

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

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 performance, 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.

Visualizing the Loss Landscape of Neural Nets

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.

Non-monotone Submodular Maximization in Exponentially Fewer Iterations

Eric Balkanski, Adam Breuer, Yaron Singer

In this paper we consider parallelization for applications whose objective can be expressed as maximizing a non-monotone submodular function under a cardinality constraint. Our main result is an algorithm whose approximation is arbitrarily close

to $1/2e$ in $O(\log^2 n)$ adaptive rounds, where n is the size of the ground set. This is an exponential speedup in parallel running time over any previously studied algorithm for constrained non-monotone submodular maximization. Beyond its provable guarantees, the algorithm performs well in practice. Specifically, experiments on traffic monitoring and personalized data summarization applications show that the algorithm finds solutions whose values are competitive with state-of-the-art algorithms while running in exponentially fewer parallel iterations.

Representation Learning for Treatment Effect Estimation from Observational Data

LiuYi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, Aidong Zhang

Estimating individual treatment effect (ITE) is a challenging problem in causal inference, due to the missing counterfactuals and the selection bias. Existing ITE estimation methods mainly focus on balancing the distributions of control and treated groups, but ignore the local similarity information that is helpful. In this paper, we propose a local similarity preserved individual treatment effect (SITE) estimation method based on deep representation learning. SITE preserves local similarity and balances data distributions simultaneously, by focusing on several hard samples in each mini-batch. Experimental results on synthetic and three real-world datasets demonstrate the advantages of the proposed SITE method, compared with the state-of-the-art ITE estimation methods.

Memory Replay GANs: Learning to Generate New Categories without Forgetting

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.

HOGWILD!-Gibbs can be PanAccurate

Constantinos Daskalakis, Nishanth Dikkala, Siddhartha Jayanti

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

DeepExposure: Learning to Expose Photos with Asynchronously Reinforced Adversarial Learning

Runsheng Yu, Wenyu Liu, Yaseen 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 simplify the algorithmic architecture, we also prove the feasibility of leveraging the discriminator as the value function. Furthermore, 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.

Data Amplification: A Unified and Competitive Approach to Property Estimation

Yi Hao, Alon Orlitsky, Ananda Theertha Suresh, Yihong Wu

Estimating properties of discrete distributions is a fundamental problem in statistical learning. We design the first unified, linear-time, competitive, property estimator that for a wide class of properties and for all underlying distributions uses just $2n$ samples to achieve the performance attained by the empirical estimator with $n\sqrt{\log n}$ samples. This provides off-the-shelf, distribution-independent, "amplification" of the amount of data available relative to common-practice estimators.

Insights on representational similarity in neural networks with canonical correlation

Ari Morcos, Maithra Raghu, Samy Bengio

Comparing different neural network representations and determining how representations evolve over time remain challenging open questions in our understanding of the function of neural networks. Comparing representations in neural networks is fundamentally difficult as the structure of representations varies greatly, even across groups of networks trained on identical tasks, and over the course of training. Here, we develop projection weighted CCA (Canonical Correlation Analysis) as a tool for understanding neural networks, building off of SVCCA, a recently proposed method (Raghu et al, 2017). We first improve the core method, showing how to differentiate between signal and noise, and then apply this technique to compare across a group of CNNs, demonstrating that networks which generalize converge to more similar representations than networks which memorize, that wider networks converge to more similar solutions than narrow networks, and that trained networks with identical topology but different learning rates converge to distinct clusters with diverse representations. We also investigate the representational dynamics of RNNs, across both training and sequential timesteps, finding that RNNs converge in a bottom-up pattern over the course of training and that the hidden state is highly variable over the course of a sequence, even when acc

ounting for linear transforms. Together, these results provide new insights into the function of CNNs and RNNs, and demonstrate the utility of using CCA to understand representations.

Efficient nonmyopic batch active search

Shali Jiang, Gustavo Malkomes, Matthew Abbott, Benjamin Moseley, Roman Garnett
Active search is a learning paradigm for actively identifying as many members of a given class as possible. A critical target scenario is high-throughput screening for scientific discovery, such as drug or materials discovery. In these settings, specialized instruments can often evaluate \emph{multiple} points simultaneously; however, all existing work on active search focuses on sequential acquisition. We bridge this gap, addressing batch active search from both the theoretical and practical perspective. We first derive the Bayesian optimal policy for this problem, then prove a lower bound on the performance gap between sequential and batch optimal policies: the ``cost of parallelization.'' We also propose novel, efficient batch policies inspired by state-of-the-art sequential policies, and develop an aggressive pruning technique that can dramatically speed up computation. We conduct thorough experiments on data from three application domains: a citation network, material science, and drug discovery, testing all proposed policies (14 total) with a wide range of batch sizes. Our results demonstrate that the empirical performance gap matches our theoretical bound, that nonmyopic policies usually significantly outperform myopic alternatives, and that diversity is an important consideration for batch policy design.

Learning Safe Policies with Expert Guidance

Jessie Huang, Fa Wu, Doina Precup, Yang Cai

We propose a framework for ensuring safe behavior of a reinforcement learning agent when the reward function may be difficult to specify. In order to do this, we rely on the existence of demonstrations from expert policies, and we provide a theoretical framework for the agent to optimize in the space of rewards consistent with its existing knowledge. We propose two methods to solve the resulting optimization: an exact ellipsoid-based method and a method in the spirit of the "follow-the-perturbed-leader" algorithm. Our experiments demonstrate the behavior of our algorithm in both discrete and continuous problems. The trained agent safely avoids states with potential negative effects while imitating the behavior of the expert in the other states.

Fast Similarity Search via Optimal Sparse Lifting

Wenye Li, Jingwei Mao, Yin Zhang, Shuguang Cui

Similarity search is a fundamental problem in computing science with various applications and has attracted significant research attention, especially in large-scale search with high dimensions. Motivated by the evidence in biological science, our work develops a novel approach for similarity search. Fundamentally different from existing methods that typically reduce the dimension of the data to lessen the computational complexity and speed up the search, our approach projects the data into an even higher-dimensional space while ensuring the sparsity of the data in the output space, with the objective of further improving precision and speed. Specifically, our approach has two key steps. Firstly, it computes the optimal sparse lifting for given input samples and increases the dimension of the data while approximately preserving their pairwise similarity. Secondly, it seeks the optimal lifting operator that best maps input samples to the optimal sparse lifting. Computationally, both steps are modeled as optimization problems that can be efficiently and effectively solved by the Frank-Wolfe algorithm. Simple as it is, our approach has reported significantly improved results in empirical evaluations, and exhibited its high potentials in solving practical problems.

Differentially Private Robust Low-Rank Approximation

Raman Arora, Vladimir Braverman, Jalaj Upadhyay

Requests for name changes in the electronic proceedings will be accepted with no

questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Evidential Deep Learning to Quantify Classification Uncertainty

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.

A Game-Theoretic Approach to Recommendation Systems with Strategic Content Providers

Omer Ben-Porat, Moshe Tennenholtz

We introduce a game-theoretic approach to the study of recommendation systems with strategic content providers. Such systems should be fair and stable. Showing that traditional approaches fail to satisfy these requirements, we propose the Shapley mediator. We show that the Shapley mediator satisfies the fairness and stability requirements, runs in linear time, and is the only economically efficient mechanism satisfying these properties.

Frequency-Domain Dynamic Pruning for Convolutional Neural Networks

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.

Adaptive Path-Integral Autoencoders: Representation Learning and Planning for Dynamical Systems

Jung-Su Ha, Young-Jin Park, Hyeok-Joo Chae, Soon-Seo Park, Han-Lim Choi

We present a representation learning algorithm that learns a low-dimensional latent dynamical system from high-dimensional sequential raw data, e.g., video. The framework builds upon recent advances in amortized inference methods that use both an inference network and a refinement procedure to output samples from a variational distribution given an observation sequence, and takes advantage of the duality between control and inference to approximately solve the intractable inference problem using the path integral control approach. The learned dynamical model can be used to predict and plan the future states; we also present the efficient planning method that exploits the learned low-dimensional latent dynamics.

Numerical experiments show that the proposed path-integral control based variational inference method leads to tighter lower bounds in statistical model learning of sequential data. Supplementary video: <https://youtu.be/xCp35crUoLQ>

Testing for Families of Distributions via the Fourier Transform

Clément L. Canonne, Ilias Diakonikolas, Alistair Stewart

We study the general problem of testing whether an unknown discrete distribution belongs to a specified family of distributions. More specifically, given a distribution family \mathcal{P} and sample access to an unknown discrete distribution D , we want to distinguish (with high probability) between the case that $D \in \mathcal{P}$ and the case that D is ϵ -far, in total variation distance, from every distribution in \mathcal{P} . This is the prototypical hypothesis testing problem that has received significant attention in statistics and, more recently, in computer science. The main contribution of this work is a simple and general testing technique that is applicable to all distribution families whose Fourier spectrum satisfies a certain approximate sparsity property. We apply our Fourier-based framework to obtain near sample-optimal and computationally efficient testers for the following fundamental distribution families: Sums of Independent Integer Random Variables (SIIRVs), Poisson Multinomial Distributions (PMDs), and Discrete Log-Concave Distributions. For the first two, ours are the first non-trivial testers in the literature, vastly generalizing previous work on testing Poisson Binomial Distributions. For the third, our tester improves on prior work in both sample and time complexity.

A Unified Framework for Extensive-Form Game Abstraction with Bounds

Christian Kroer, Tuomas Sandholm

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Model-Agnostic Private Learning

Raef Bassily, Om Thakkar, Abhradeep Guha Thakurta

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Towards Text Generation with Adversarially Learned Neural Outlines

Sandeep Subramanian, Sai Rajeswar Mudumba, Alessandro Sordani, Adam Trischler, Aaron C. 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.

FastGRNN: A Fast, Accurate, Stable and Tiny Kilobyte Sized Gated Recurrent Neural Network

Aditya Kusupati, Manish Singh, Kush Bhatia, Ashish Kumar, Prateek Jain, Manik Varma

This paper develops the FastRNN and FastGRNN algorithms to address the twin RNN

limitations of inaccurate training and inefficient prediction. Previous approaches have improved accuracy at the expense of prediction costs making them infeasible for resource-constrained and real-time applications. Unitary RNNs have increased accuracy somewhat by restricting the range of the state transition matrix's singular values but have also increased the model size as they require a larger number of hidden units to make up for the loss in expressive power. Gated RNNs have obtained state-of-the-art accuracies by adding extra parameters thereby resulting in even larger models. FastRNN addresses these limitations by adding a residual connection that does not constrain the range of the singular values explicitly and has only two extra scalar parameters. FastGRNN then extends the residual connection to a gate by reusing the RNN matrices to match state-of-the-art gated RNN accuracies but with a 2-4x smaller model. Enforcing FastGRNN's matrices to be low-rank, sparse and quantized resulted in accurate models that could be up to 35x smaller than leading gated and unitary RNNs. This allowed FastGRNN to accurately recognize the "Hey Cortana" wakeword with a 1 KB model and to be deployed on severely resource-constrained IoT microcontrollers too tiny to store other RNN models. FastGRNN's code is available at (<https://github.com/Microsoft/EdgeML/>).

Bipartite Stochastic Block Models with Tiny Clusters

Stefan Neumann

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Conditional Adversarial Domain Adaptation

Mingsheng Long, ZHANGJIE CAO, Jianmin Wang, Michael I. 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.

Stochastic Expectation Maximization with Variance Reduction

Jianfei Chen, Jun Zhu, Yee Whye Teh, Tong Zhang

Expectation-Maximization (EM) is a popular tool for learning latent variable models, but the vanilla batch EM does not scale to large data sets because the whole data set is needed at every E-step. Stochastic Expectation Maximization (sEM) reduces the cost of E-step by stochastic approximation. However, sEM has a slower asymptotic convergence rate than batch EM, and requires a decreasing sequence of step sizes, which is difficult to tune. In this paper, we propose a variance reduced stochastic EM (sEM-vr) algorithm inspired by variance reduced stochastic gradient descent algorithms. We show that sEM-vr has the same exponential asymptotic convergence rate as batch EM. Moreover, sEM-vr only requires a constant step size to achieve this rate, which alleviates the burden of parameter tuning. We compare sEM-vr with batch EM, sEM and other algorithms on Gaussian mixture models and probabilistic latent semantic analysis, and sEM-vr converges significantly faster than these baselines.

Bayesian Nonparametric Spectral Estimation

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 optimized 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.

A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

Kimin Lee, Kibok Lee, Honglak Lee, Jinwoo Shin

Detecting test samples drawn sufficiently far away from the training distribution statistically or adversarially is a fundamental requirement for deploying a good classifier in many real-world machine learning applications. However, deep neural networks with the softmax classifier are known to produce highly overconfident posterior distributions even for such abnormal samples. In this paper, we propose a simple yet effective method for detecting any abnormal samples, which is applicable to any pre-trained softmax neural classifier. We obtain the class conditional Gaussian distributions with respect to (low- and upper-level) features of the deep models under Gaussian discriminant analysis, which result in a confidence score based on the Mahalanobis distance. While most prior methods have been evaluated for detecting either out-of-distribution or adversarial samples, but not both, the proposed method achieves the state-of-the-art performances for both cases in our experiments. Moreover, we found that our proposed method is more robust in harsh cases, e.g., when the training dataset has noisy labels or small number of samples. Finally, we show that the proposed method enjoys broader usage by applying it to class-incremental learning: whenever out-of-distribution samples are detected, our classification rule can incorporate new classes well without further training deep models.

Breaking the Span Assumption Yields Fast Finite-Sum Minimization

Robert Hannah, Yanli Liu, Daniel O'Connor, Wotao Yin

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Differential Privacy for Growing Databases

Rachel Cummings, Sara Krehbiel, Kevin A. Lai, Uthaipon Tantipongpipat

The large majority of differentially private algorithms focus on the static setting, where queries are made on an unchanging database. This is unsuitable for the myriad applications involving databases that grow over time. To address this gap in the literature, we consider the dynamic setting, in which new data arrive over time. Previous results in this setting have been limited to answering a single non-adaptive query repeatedly as the database grows. In contrast, we provide tools for richer and more adaptive analysis of growing databases. Our first contribution is a novel modification of the private multiplicative weights algorithm, which provides accurate analysis of exponentially many adaptive linear queries (an expressive query class including all counting queries) for a static database. Our modification maintains the accuracy guarantee of the static setting even as the database grows without bound. Our second contribution is a set of general results which show that many other private and accurate algorithms can be immediately extended to the dynamic setting by rerunning them at appropriate points of data growth with minimal loss of accuracy, even when data growth is unbounded


.

Learning Pipelines with Limited Data and Domain Knowledge: A Study in Parsing Physics Problems

Mrinmaya Sachan, Kumar Avinava Dubey, Tom M. Mitchell, Dan Roth, Eric P. Xing

As machine learning becomes more widely used in practice, we need new methods to build complex intelligent systems that integrate learning with existing software, and with domain knowledge encoded as rules. As a case study, we present such a system that learns to parse Newtonian physics problems in textbooks. This system, Nuts&Bolts, learns a pipeline process that incorporates existing code, pre-learned machine learning models, and human engineered rules. It jointly trains the entire pipeline to prevent propagation of errors, using a combination of labeled and unlabelled data. Our approach achieves a good performance on the parsing task, outperforming the simple pipeline and its variants. Finally, we also show how Nuts&Bolts can be used to achieve improvements on a relation extraction task and on the end task of answering Newtonian physics problems.

Theoretical guarantees for EM under misspecified Gaussian mixture models

Raaz Dwivedi,  H, Koulik Khamaru, Martin J. Wainwright, Michael I. Jordan

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Online Improper Learning with an Approximation Oracle

Elad Hazan, Wei Hu, Yuanzhi Li, Zhiyuan Li

We study the following question: given an efficient approximation algorithm for an optimization problem, can we learn efficiently in the same setting? We give a formal affirmative answer to this question in the form of a reduction from online learning to offline approximate optimization using an efficient algorithm that guarantees near optimal regret. The algorithm is efficient in terms of the number of oracle calls to a given approximation oracle - it makes only logarithmically many such calls per iteration. This resolves an open question by Kalai and Vempala, and by Garber. Furthermore, our result applies to the more general improper learning problems.

Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, Kevin Gimpel

The growing importance of massive datasets with the advent of deep learning makes robustness to label noise a critical property for classifiers to have. Sources of label noise include automatic labeling for large datasets, non-expert labeling, and label corruption by data poisoning adversaries. In the latter case, corruptions may be arbitrarily bad, even so bad that a classifier predicts the wrong labels with high confidence. To protect against such sources of noise, we leverage the fact that a small set of clean labels is often easy to procure. We demonstrate that robustness to label noise up to severe strengths can be achieved by using a set of trusted data with clean labels, and propose a loss correction that utilizes trusted examples in a data-efficient manner to mitigate the effects of label noise on deep neural network classifiers. Across vision and natural language processing tasks, we experiment with various label noises at several strengths, and show that our method significantly outperforms existing methods.

Multi-Task Zipping via Layer-wise Neuron Sharing

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

ing (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.

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

Masking: A New Perspective of Noisy Supervision

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.

Learning to Multitask

Yu Zhang, Ying Wei, Qiang Yang

Multitask learning has shown promising performance in many applications and many multitask models have been proposed. In order to identify an effective multitask model for a given multitask problem, we propose a learning framework called Learning to MultiTask (L2MT). To achieve the goal, L2MT exploits historical multitask experience which is organized as a training set consisting of several tuples, each of which contains a multitask problem with multiple tasks, a multitask model, and the relative test error. Based on such training set, L2MT first uses a proposed layerwise graph neural network to learn task embeddings for all the tasks in a multitask problem and then learns an estimation function to estimate the relative test error based on task embeddings and the representation of the multitask model based on a unified formulation. Given a new multitask problem, the estimation function is used to identify a suitable multitask model. Experiments on benchmark datasets show the effectiveness of the proposed L2MT framework.

Thwarting Adversarial Examples: An \mathcal{L}_0 -Robust Sparse Fourier Transform

Mitali Bafna, Jack Murtagh, Nikhil Vyas

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Constant Regret, Generalized Mixability, and Mirror Descent

Zakaria Mhammedi, Robert C. Williamson

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Dual Principal Component Pursuit: Improved Analysis and Efficient Algorithms

Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, René Vidal, Manolis Tsakiris

Recent methods for learning a linear subspace from data corrupted by outliers are based on convex L1 and nuclear norm optimization and require the dimension of the subspace and the number of outliers to be sufficiently small [27]. In sharp contrast, the recently proposed Dual Principal Component Pursuit (DPCP) method [22] can provably handle subspaces of high dimension by solving a non-convex L1 optimization problem on the sphere. However, its geometric analysis is based on quantities that are difficult to interpret and are not amenable to statistical analysis. In this paper we provide a refined geometric analysis and a new statistical analysis that show that DPCP can tolerate as many outliers as the square of the number of inliers, thus improving upon other provably correct robust PCA methods. We also propose a scalable Projected Sub-Gradient Descent method (DPCP-PSGD) for solving the DPCP problem and show it admits linear convergence even though the underlying optimization problem is non-convex and non-smooth. Experiments on road plane detection from 3D point cloud data demonstrate that DPCP-PSGD can be more efficient than the traditional RANSAC algorithm, which is one of the most popular methods for such computer vision applications.

Generative modeling for protein structures

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.

Found Graph Data and Planted Vertex Covers

Austin R. Benson, Jon Kleinberg

A typical way in which network data is recorded is to measure all interactions involving a specified set of core nodes, which produces a graph containing this core together with a potentially larger set of fringe nodes that link to the core. Interactions between nodes in the fringe, however, are not present in the resulting graph data. For example, a phone service provider may only record calls in which at least one of the participants is a customer; this can include calls between a customer and a non-customer, but not between pairs of non-customers. Knowledge of which nodes belong to the core is crucial for interpreting the dataset, but this metadata is unavailable in many cases, either because it has been lost due to difficulties in data provenance, or because the network consists of "fo

und data" obtained in settings such as counter-surveillance. This leads to an algorithmic problem of recovering the core set. Since the core is a vertex cover, we essentially have a planted vertex cover problem, but with an arbitrary underlying graph. We develop a framework for analyzing this planted vertex cover problem, based on the theory of fixed-parameter tractability, together with algorithms for recovering the core. Our algorithms are fast, simple to implement, and outperform several baselines based on core-periphery structure on various real-world datasets.

Fast Estimation of Causal Interactions using Wold Processes

Flavio Figueiredo, Guilherme Resende Borges, Pedro O.S. Vaz de Melo, Renato Assunção

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Relating Leverage Scores and Density using Regularized Christoffel Functions

Edouard Pauwels, Francis Bach, Jean-Philippe Vert

Statistical leverage scores emerged as a fundamental tool for matrix sketching and column sampling with applications to low rank approximation, regression, random feature learning and quadrature. Yet, the very nature of this quantity is barely understood. Borrowing ideas from the orthogonal polynomial literature, we introduce the regularized Christoffel function associated to a positive definite kernel. This uncovers a variational formulation for leverage scores for kernel methods and allows to elucidate their relationships with the chosen kernel as well as population density. Our main result quantitatively describes a decreasing relation between leverage score and population density for a broad class of kernels on Euclidean spaces. Numerical simulations support our findings.

Online Adaptive Methods, Universality and Acceleration

Kfir Y. Levy, Alp Yurtsever, Volkan Cevher

We present a novel method for convex unconstrained optimization that, without any modifications ensures: (1) accelerated convergence rate for smooth objectives, (2) standard convergence rate in the general (non-smooth) setting, and (3) standard convergence rate in the stochastic optimization setting.

To the best of our knowledge, this is the first method that simultaneously applies to all of the above settings.

At the heart of our method is an adaptive learning rate rule that employs importance weights, in the spirit of adaptive online learning algorithms [duchi2011adaptive, levy2017online], combined with an update that linearly couples two sequences, in the spirit of [AllenOrecchia2017]. An empirical examination of our method demonstrates its applicability to the above mentioned scenarios and corroborates our theoretical findings.

Reparameterization Gradient for Non-differentiable Models

Wonyeol Lee, Hangeol Yu, Hongseok Yang

We present a new algorithm for stochastic variational inference that targets at models with non-differentiable densities. One of the key challenges in stochastic variational inference is to come up with a low-variance estimator of the gradient of a variational objective. We tackle the challenge by generalizing the reparameterization trick, one of the most effective techniques for addressing the variance issue for differentiable models, so that the trick works for non-differentiable models as well. Our algorithm splits the space of latent variables into regions where the density of the variables is differentiable, and their boundaries where the density may fail to be differentiable. For each differentiable region, the algorithm applies the standard reparameterization trick and estimates the gradient restricted to the region. For each potentially non-differentiable boundary, it uses a form of manifold sampling and computes the direction for variational parameters that, if followed, would increase the boundary's contribution to

the variational objective. The sum of all the estimates becomes the gradient estimate of our algorithm. Our estimator enjoys the reduced variance of the reparameterization gradient while remaining unbiased even for non-differentiable models. The experiments with our preliminary implementation confirm the benefit of reduced variance and unbiasedness.

Improving Exploration in Evolution Strategies for Deep Reinforcement Learning via a Population of Novelty-Seeking Agents

Edoardo Conti, Vashisht Madhavan, Felipe Petroski Such, Joel Lehman, Kenneth Stanley, Jeff Clune

Evolution strategies (ES) are a family of black-box optimization algorithms able to train deep neural networks roughly as well as Q-learning and policy gradient methods on challenging deep reinforcement learning (RL) problems, but are much faster (e.g. hours vs. days) because they parallelize better. However, many RL problems require directed exploration because they have reward functions that are sparse or deceptive (i.e. contain local optima), and it is unknown how to encourage such exploration with ES. Here we show that algorithms that have been invented to promote directed exploration in small-scale evolved neural networks via populations of exploring agents, specifically novelty search (NS) and quality diversity (QD) algorithms, can be hybridized with ES to improve its performance on sparse or deceptive deep RL tasks, while retaining scalability. Our experiments confirm that the resultant new algorithms, NS-ES and two QD algorithms, NSR-ES and NSRA-ES, avoid local optima encountered by ES to achieve higher performance on Atari and simulated robots learning to walk around a deceptive trap. This paper thus introduces a family of fast, scalable algorithms for reinforcement learning that are capable of directed exploration. It also adds this new family of exploration algorithms to the RL toolbox and raises the interesting possibility that analogous algorithms with multiple simultaneous paths of exploration might also combine well with existing RL algorithms outside ES.

Optimistic optimization of a Brownian

Jean-Bastien Grill, Michal Valko, Remi Munos

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Generalizing Tree Probability Estimation via Bayesian Networks

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.

Safe Active Learning for Time-Series Modeling with Gaussian Processes

Christoph Zimmer, Mona Meister, Duy Nguyen-Tuong

Learning time-series models is useful for many applications, such as simulation and forecasting. In this study, we consider the problem of actively learning time-series models while taking given safety constraints into account. For time-series modeling we employ a Gaussian process with a nonlinear exogenous input structure. The proposed approach generates data appropriate for time series model learning, i.e. input and output trajectories, by dynamically exploring the input space. The approach parametrizes the input trajectory as consecutive trajectory sections, which are determined stepwise given safety requirements and past observa-

tions. We analyze the proposed algorithm and evaluate it empirically on a technical application. The results show the effectiveness of our approach in a realistic technical use case.

Computing Kantorovich-Wasserstein Distances on d -dimensional histograms using $(d+1)$ -partite graphs

Gennaro Auricchio, Federico Bassetti, Stefano Gualandi, Marco Veneroni

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Simple Embedding for Link Prediction in Knowledge Graphs

Seyed Mehran Kazemi, David Poole

Knowledge graphs contain knowledge about the world and provide a structured representation of this knowledge. Current knowledge graphs contain only a small subset of what is true in the world. Link prediction approaches aim at predicting new links for a knowledge graph given the existing links among the entities. Tensor factorization approaches have proved promising for such link prediction problems. Proposed in 1927, Canonical Polyadic (CP) decomposition is among the first tensor factorization approaches. CP generally performs poorly for link prediction as it learns two independent embedding vectors for each entity, whereas they are really tied. We present a simple enhancement of CP (which we call Simple) to allow the two embeddings of each entity to be learned dependently. The complexity of Simple grows linearly with the size of embeddings. The embeddings learned through Simple are interpretable, and certain types of background knowledge can be incorporated into these embeddings through weight tying.

We prove Simple is fully expressive and derive a bound on the size of its embeddings for full expressivity.

We show empirically that, despite its simplicity, Simple outperforms several state-of-the-art tensor factorization techniques.

Simple's code is available on GitHub at <https://github.com/Mehran-k/Simple>.

Bounded-Loss Private Prediction Markets

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.

Statistical mechanics of low-rank tensor decomposition

Jonathan Kadmon, Surya Ganguli

Often, large, high dimensional datasets collected across multiple modalities can be organized as a higher order tensor. Low-rank tensor decomposition then arises as a powerful and widely used tool to discover simple low dimensional structures underlying such data. However, we currently lack a theoretical understanding of the algorithmic behavior of low-rank tensor decompositions. We derive Bayesian approximate message passing (AMP) algorithms for recovering arbitrarily shaped low-rank tensors buried within noise, and we employ dynamic mean field theory to precisely characterize their performance. Our theory reveals the existence of phase transitions between easy, hard and impossible inference regimes, and displays an excellent match with simulations.

Moreover, it reveals several qualitative surprises compared to the behavior of symmetric, cubic tensor decomposition. Finally, we compare our AMP algorithm to the most commonly used algorithm, alternating least squares (ALS), and demonstrate that AMP significantly outperforms ALS in the presence of noise.

A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization

Cedric Jozs, Yi Ouyang, Richard Zhang, Javad Lavaei, Somayeh Sojoudi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

A Structured Prediction Approach for Label Ranking

Anna Korba, Alexandre Garcia, Florence d'Alché-Buc

We propose to solve a label ranking problem as a structured output regression task. In this view, we adopt a least square surrogate loss

approach that solves a supervised learning problem in two steps:

a regression step in a well-chosen feature space and a pre-image (or decoding) step. We use specific feature maps/embeddings for ranking data, which convert any ranking/permutation into a vector representation. These embeddings are all well-tailored for our approach, either by resulting in consistent estimators, or by solving trivially the pre-image problem which is often the bottleneck in structured prediction. Their extension to the case of incomplete or partial rankings is also discussed. Finally, we provide empirical results on synthetic and real-world datasets showing the relevance of our method.

Geometrically Coupled Monte Carlo Sampling

Mark Rowland, Krzysztof M. Choromanski, François Chalus, Aldo Pacchiano, Tamas Szepesvári, Richard E. Turner, Adrian Weller

Monte Carlo sampling in high-dimensional, low-sample settings is important in many machine learning tasks. We improve current methods for sampling in Euclidean spaces by avoiding independence, and instead consider ways to couple samples. We show fundamental connections to optimal transport theory, leading to novel sampling algorithms, and providing new theoretical grounding for existing strategies. We compare our new strategies against prior methods for improving sample efficiency, including QMC, by studying discrepancy. We explore our findings empirically, and observe benefits of our sampling schemes for reinforcement learning and generative modelling.

The Lingering of Gradients: How to Reuse Gradients Over Time

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

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Submodular Maximization via Gradient Ascent: The Case of Deep Submodular Functions

Wenruo Bai, William Stafford Noble, Jeff A. Bilmes

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Sparsified SGD with Memory

Sebastian U. Stich, Jean-Baptiste Cordonnier, Martin Jaggi

Huge scale machine learning problems are nowadays tackled by distributed optimization algorithms, i.e. algorithms that leverage the compute power of many device

s for training. The communication overhead is a key bottleneck that hinders perfect scalability. Various recent works proposed to use quantization or sparsification techniques to reduce the amount of data that needs to be communicated, for instance by only sending the most significant entries of the stochastic gradient (top-k sparsification). Whilst such schemes showed very promising performance in practice, they have eluded theoretical analysis so far.

Convergence of Cubic Regularization for Nonconvex Optimization under KL Property
Yi Zhou, Zhe Wang, Yingbin Liang

Cubic-regularized Newton's method (CR) is a popular algorithm that guarantees to produce a second-order stationary solution for solving nonconvex optimization problems. However, existing understandings of convergence rate of CR are conditioned on special types of geometrical properties of the objective function. In this paper, we explore the asymptotic convergence rate of CR by exploiting the ubiquitous Kurdyka-Lojasiewicz (KL) property of the nonconvex objective functions. In specific, we characterize the asymptotic convergence rate of various types of optimality measures for CR including function value gap, variable distance gap, gradient norm and least eigenvalue of the Hessian matrix. Our results fully characterize the diverse convergence behaviors of these optimality measures in the full parameter regime of the KL property. Moreover, we show that the obtained asymptotic convergence rates of CR are order-wise faster than those of first-order gradient descent algorithms under the KL property.

Model Agnostic Supervised Local Explanations

Gregory Plumb, Denali Molitor, Ameet S. 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.

Mental Sampling in Multimodal Representations

Jianqiao Zhu, Adam Sanborn, Nick Chater

Both resources in the natural environment and concepts in a semantic space are distributed "patchily", with large gaps in between the patches. To describe people's internal and external foraging behavior, various random walk models have been proposed. In particular, internal foraging has been modeled as sampling: in order to gather relevant information for making a decision, people draw samples from a mental representation using random-walk algorithms such as Markov chain Monte Carlo (MCMC). However, two common empirical observations argue against people using simple sampling algorithms such as MCMC for internal foraging. First, the distance between samples is often best described by a Levy flight distribution: the probability of the distance between two successive locations follows a power-law on the distances. Second, humans and other animals produce long-range, slowly decaying autocorrelations characterized as $1/f$ -like fluctuations, instead of the $1/f^2$ fluctuations produced by random walks. We propose that mental sampling is not done by simple MCMC, but is instead adapted to multimodal representations and is implemented by Metropolis-coupled Markov chain Monte Carlo (MC3), one

of the first algorithms developed for sampling from multimodal distributions. MC3 involves running multiple Markov chains in parallel but with target distributions of different temperatures, and it swaps the states of the chains whenever a better location is found. Heated chains more readily traverse valleys in the probability landscape to propose moves to far-away peaks, while the colder chains make the local steps that explore the current peak or patch. We show that MC3 generates distances between successive samples that follow a Levy flight distribution and produce $1/f$ -like autocorrelations, providing a single mechanistic account of these two puzzling empirical phenomena of internal foraging.

Nonparametric learning from Bayesian models with randomized objective functions
Simon Lyddon, Stephen Walker, Chris C. Holmes

Bayesian learning is built on an assumption that the model space contains a true reflection of the data generating mechanism. This assumption is problematic, particularly in complex data environments. Here we present a Bayesian nonparametric approach to learning that makes use of statistical models, but does not assume that the model is true. Our approach has provably better properties than using a parametric model and admits a Monte Carlo sampling scheme that can afford massive scalability on modern computer architectures. The model-based aspect of learning is particularly attractive for regularizing nonparametric inference when the sample size is small, and also for correcting approximate approaches such as variational Bayes (VB). We demonstrate the approach on a number of examples including VB classifiers and Bayesian random forests.

On the Dimensionality of Word Embedding

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.

Large Scale computation of Means and Clusters for Persistence Diagrams using Optimal Transport

Theo Lacombe, Marco Cuturi, Steve OUDOT

Persistence diagrams (PDs) are now routinely used to summarize the underlying topology of complex data. Despite several appealing properties, incorporating PDs in learning pipelines can be challenging because their natural geometry is not Hilbertian. Indeed, this was recently exemplified in a string of papers which show that the simple task of averaging a few PDs can be computationally prohibitive. We propose in this article a tractable framework to carry out standard tasks on PDs at scale, notably evaluating distances, estimating barycenters and performing clustering. This framework builds upon a reformulation of PD metrics as optimal transport (OT) problems. Doing so, we can exploit recent computational advances: the OT problem on a planar grid, when regularized with entropy, is convex and can be solved in linear time using the Sinkhorn algorithm and convolutions. This results in scalable computations that can stream on GPUs. We demonstrate the efficiency of our approach by carrying out clustering with diagrams metrics on several thousands of PDs, a scale never seen before in the literature.

Probabilistic Matrix Factorization for Automated Machine Learning

Nicolo Fusi, Rishit Sheth, Melih Elibol

In order to achieve state-of-the-art performance, modern machine learning techniques require careful data pre-processing and hyperparameter tuning. Moreover, gi

ven the ever increasing number of machine learning models being developed, model selection is becoming increasingly important. Automating the selection and tuning of machine learning pipelines, which can include different data pre-processing methods and machine learning models, has long been one of the goals of the machine learning community.

In this paper, we propose to solve this meta-learning task by combining ideas from collaborative filtering and Bayesian optimization. Specifically, we use a probabilistic matrix factorization model to transfer knowledge across experiments performed in hundreds of different datasets and use an acquisition function to guide the exploration of the space of possible ML pipelines. In our experiments, we show that our approach quickly identifies high-performing pipelines across a wide range of datasets, significantly outperforming the current state-of-the-art.

REFUEL: Exploring Sparse Features in Deep Reinforcement Learning for Fast Disease Diagnosis

Yu-Shao Peng, Kai-Fu Tang, Hsuan-Tien Lin, Edward Chang

This paper proposes REFUEL, a reinforcement learning method with two techniques: {\em reward shaping} and {\em feature rebuilding}, to improve the performance of online symptom checking for disease diagnosis. Reward shaping can guide the search of policy towards better directions. Feature rebuilding can guide the agent to learn correlations between features. Together, they can find symptom queries that can yield positive responses from a patient with high probability. Experimental results justify that the two techniques in REFUEL allows the symptom checker to identify the disease more rapidly and accurately.

Cluster Variational Approximations for Structure Learning of Continuous-Time Bayesian Networks from Incomplete Data

Dominik Linzner, Heinz Koeppel

Continuous-time Bayesian networks (CTBNs) constitute a general and powerful framework for modeling continuous-time stochastic processes on networks. This makes them particularly attractive for learning the directed structures among interacting entities. However, if the available data is incomplete, one needs to simulate the prohibitively complex CTBN dynamics. Existing approximation techniques, such as sampling and low-order variational methods, either scale unfavorably in system size, or are unsatisfactory in terms of accuracy. Inspired by recent advances in statistical physics, we present a new approximation scheme based on cluster-variational methods that significantly improves upon existing variational approximations. We can analytically marginalize the parameters of the approximate CTBN, as these are of secondary importance for structure learning. This recovers a scalable scheme for direct structure learning from incomplete and noisy time-series data. Our approach outperforms existing methods in terms of scalability.

Norm-Ranging LSH for Maximum Inner Product Search

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.

Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions

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.

Dirichlet-based Gaussian Processes for Large-scale Calibrated Classification
Dimitrios Miliotis, Raffaello Camoriano, Pietro Michiardi, Lorenzo Rosasco, Maurizio Filippone

This paper studies the problem of deriving fast and accurate classification algorithms with uncertainty quantification. Gaussian process classification provides a principled approach, but the corresponding computational burden is hardly sustainable in large-scale problems and devising efficient alternatives is a challenge. In this work, we investigate if and how Gaussian process regression directly applied to classification labels can be used to tackle this question. While in this case training is remarkably faster, predictions need to be calibrated for classification and uncertainty estimation. To this aim, we propose a novel regression approach where the labels are obtained through the interpretation of classification labels as the coefficients of a degenerate Dirichlet distribution. Extensive experimental results show that the proposed approach provides essentially the same accuracy and uncertainty quantification as Gaussian process classification while requiring only a fraction of computational resources.

Latent Alignment and Variational Attention

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.

The Pessimistic Limits and Possibilities of Margin-based Losses in Semi-supervised Learning

Jesse Krijthe, Marco Loog

Consider a classification problem where we have both labeled and unlabeled data

available. We show that for linear classifiers defined by convex margin-based surrogate losses that are decreasing, it is impossible to construct *any* semi-supervised approach that is able to guarantee an improvement over the supervised classifier measured by this surrogate loss on the labeled and unlabeled data. For convex margin-based loss functions that also increase, we demonstrate safe improvements *are* possible.

Porcupine Neural Networks: Approximating Neural Network Landscapes

Soheil Feizi, Hamid Javadi, Jesse Zhang, David Tse

Neural networks have been used prominently in several machine learning and statistics applications. In general, the underlying optimization of neural networks is non-convex which makes analyzing their performance challenging. In this paper, we take another approach to this problem by constraining the network such that the corresponding optimization landscape has good theoretical properties without significantly compromising performance. In particular, for two-layer neural networks we introduce Porcupine Neural Networks (PNNs) whose weight vectors are constrained to lie over a finite set of lines. We show that most local optima of PNN optimizations are global while we have a characterization of regions where bad local optimizers may exist. Moreover, our theoretical and empirical results suggest that an unconstrained neural network can be approximated using a polynomially-large PNN.

On the Local Hessian in Back-propagation

Huishuai Zhang, Wei Chen, Tie-Yan Liu

Back-propagation (BP) is the foundation for successfully training deep neural networks. However, BP sometimes has difficulties in propagating a learning signal deep enough effectively, e.g., the vanishing gradient phenomenon. Meanwhile, BP often works well when combining with "designing tricks" like orthogonal initialization, batch normalization and skip connection. There is no clear understanding on what is essential to the efficiency of BP. In this paper, we take one step towards clarifying this problem. We view BP as a solution of back-matching propagation which minimizes a sequence of back-matching losses each corresponding to one block of the network. We study the Hessian of the local back-matching loss (local Hessian) and connect it to the efficiency of BP. It turns out that those designing tricks facilitate BP by improving the spectrum of local Hessian. In addition, we can utilize the local Hessian to balance the training pace of each block and design new training algorithms. Based on a scalar approximation of local Hessian, we propose a scale-amended SGD algorithm. We apply it to train neural networks with batch normalization, and achieve favorable results over vanilla SGD. This corroborates the importance of local Hessian from another side.

Infinite-Horizon Gaussian Processes

Arno Solin, James Hensman, Richard E. Turner

Gaussian processes provide a flexible framework for forecasting, removing noise, and interpreting long temporal datasets. State space modelling (Kalman filtering) enables these non-parametric models to be deployed on long datasets by reducing the complexity to linear in the number of data points. The complexity is still cubic in the state dimension m which is an impediment to practical application. In certain special cases (Gaussian likelihood, regular spacing) the GP posterior will reach a steady posterior state when the data are very long. We leverage this and formulate an inference scheme for GPs with general likelihoods, where inference is based on single-sweep EP (assumed density filtering). The infinite-horizon model tackles the cubic cost in the state dimensionality and reduces the cost in the state dimension m to $O(m^2)$ per data point. The model is extended to online-learning of hyperparameters. We show examples for large finite-length modelling problems, and present how the method runs in real-time on a smartphone on a continuous data stream updated at 100 Hz.

Constrained Graph Variational Autoencoders for Molecule Design

Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, Alexander Gaunt

Graphs are ubiquitous data structures for representing interactions between entities. With an emphasis on applications in chemistry, we explore the task of learning to generate graphs that conform to a distribution observed in training data. We propose a variational autoencoder model in which both encoder and decoder are graph-structured. Our decoder assumes a sequential ordering of graph extension steps and we discuss and analyze design choices that mitigate the potential downsides of this linearization. Experiments compare our approach with a wide range of baselines on the molecule generation task and show that our method is successful at matching the statistics of the original dataset on semantically important metrics. Furthermore, we show that by using appropriate shaping of the latent space, our model allows us to design molecules that are (locally) optimal in desired properties.

Hardware Conditioned Policies for Multi-Robot Transfer Learning

Tao Chen, Adithyavairavan Murali, Abhinav Gupta

Deep reinforcement learning could be used to learn dexterous robotic policies but it is challenging to transfer them to new robots with vastly different hardware properties. It is also prohibitively expensive to learn a new policy from scratch for each robot hardware due to the high sample complexity of modern state-of-the-art algorithms. We propose a novel approach called Hardware Conditioned Policies where we train a universal policy conditioned on a vector representation of robot hardware. We considered robots in simulation with varied dynamics, kinematic structure, kinematic lengths and degrees-of-freedom. First, we use the kinematic structure directly as the hardware encoding and show great zero-shot transfer to completely novel robots not seen during training. For robots with lower zero-shot success rate, we also demonstrate that fine-tuning the policy network is significantly more sample-efficient than training a model from scratch. In tasks where knowing the agent dynamics is important for success, we learn an embedding for robot hardware and show that policies conditioned on the encoding of hardware tend to generalize and transfer well. Videos of experiments are available at: <https://sites.google.com/view/robot-transfer-hcp>.

Learning without the Phase: Regularized PhaseMax Achieves Optimal Sample Complexity

Fariborz Salehi, Ehsan Abbasi, Babak Hassibi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Learning Disentangled Joint Continuous and Discrete Representations

Emilien Dupont

We present a framework for learning disentangled and interpretable jointly continuous and discrete representations in an unsupervised manner. By augmenting the continuous latent distribution of variational autoencoders with a relaxed discrete distribution and controlling the amount of information encoded in each latent unit, we show how continuous and categorical factors of variation can be discovered automatically from data. Experiments show that the framework disentangles continuous and discrete generative factors on various datasets and outperforms current disentangling methods when a discrete generative factor is prominent.

Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples

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

ts 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.

Learning To Learn Around A Common Mean

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, Massimiliano Pontil

The problem of learning-to-learn (LTL) or meta-learning is gaining increasing attention due to recent empirical evidence of its effectiveness in applications. The goal addressed in LTL is to select an algorithm that works well on tasks sampled from a meta-distribution. In this work, we consider the family of algorithms given by a variant of Ridge Regression, in which the regularizer is the square distance to an unknown mean vector. We show that, in this setting, the LTL problem can be reformulated as a Least Squares (LS) problem and we exploit a novel meta-algorithm to efficiently solve it. At each iteration the meta-algorithm processes only one dataset. Specifically, it firstly estimates the stochastic LS objective function, by splitting this dataset into two subsets used to train and test the inner algorithm, respectively. Secondly, it performs a stochastic gradient step with the estimated value. Under specific assumptions, we present a bound for the generalization error of our meta-algorithm, which suggests the right splitting parameter to choose. When the hyper-parameters of the problem are fixed, this bound is consistent as the number of tasks grows, even if the sample size is kept constant. Preliminary experiments confirm our theoretical findings, highlighting the advantage of our approach, with respect to independent task learning.

Recurrent Relational Networks

Rasmus Palm, Ulrich Paquet, Ole Winther

This paper is concerned with learning to solve tasks that require a chain of interdependent 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 network, 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 relat

ional

networks. Finally, we show how recurrent relational networks can learn to solve Sudoku puzzles from supervised training data, a challenging task requiring upwads

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.

Experimental Design for Cost-Aware Learning of Causal Graphs

Erik Lindgren, Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath

We consider the minimum cost intervention design problem: Given the essential graph of a causal graph and a cost to intervene on a variable, identify the set of interventions with minimum total cost that can learn any causal graph with the given essential graph. We first show that this problem is NP-hard. We then prove that we can achieve a constant factor approximation to this problem with a greedy algorithm. We then constrain the sparsity of each intervention. We develop an algorithm that returns an intervention design that is nearly optimal in terms of size for sparse graphs with sparse interventions and we discuss how to use it when there are costs on the vertices.

Reinforcement Learning with Multiple Experts: A Bayesian Model Combination Approach

Michael Gimelfarb, Scott Sanner, Chi-Guhn Lee

Potential based reward shaping is a powerful technique for accelerating convergence of reinforcement learning algorithms. Typically, such information includes an estimate of the optimal value function and is often provided by a human expert or other sources of domain knowledge. However, this information is often biased or inaccurate and can mislead many reinforcement learning algorithms. In this paper, we apply Bayesian Model Combination with multiple experts in a way that learns to trust a good combination of experts as training progresses. This approach is both computationally efficient and general, and is shown numerically to improve convergence across discrete and continuous domains and different reinforcement learning algorithms.

Differentiable MPC for End-to-end Planning and Control

Brandon Amos, Ivan Jimenez, Jacob Sacks, Byron Boots, J. Zico Kolter

We present foundations for using Model Predictive Control (MPC) as a differentiable policy class for reinforcement learning. This provides one way of leveraging and combining the advantages of model-free and model-based approaches. Specifically, we differentiate through MPC by using the KKT conditions of the convex approximation at a fixed point of the controller. Using this strategy, we are able to learn the cost and dynamics of a controller via end-to-end learning. Our experiments focus on imitation learning in the pendulum and cartpole domains, where we learn the cost and dynamics terms of an MPC policy class. We show that our MPC policies are significantly more data-efficient than a generic neural network and that our method is superior to traditional system identification in a setting where the expert is unrealizable.

Zeroth-Order Stochastic Variance Reduction for Nonconvex Optimization

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, Lisa Amini

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, Yinyu Ye

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors

ors prior to requesting a name change in the electronic proceedings.

Algebraic tests of general Gaussian latent tree models

Dennis Leung, Mathias Drton

We consider general Gaussian latent tree models in which the observed variables are not restricted to be leaves of the tree. Extending related recent work, we give a full semi-algebraic description of the set of covariance matrices of any such model. In other words, we find polynomial constraints that characterize when a matrix is the covariance matrix of a distribution in a given latent tree model. However, leveraging these constraints to test a given such model is often complicated by the number of constraints being large and by singularities of individual polynomials, which may invalidate standard approximations to relevant probability distributions. Illustrating with the star tree, we propose a new testing methodology that circumvents singularity issues by trading off some statistical estimation efficiency and handles cases with many constraints through recent advances on Gaussian approximation for maxima of sums of high-dimensional random vectors. Our test avoids the need to maximize the possibly multimodal likelihood function of such models and is applicable to models with larger number of variables. These points are illustrated in numerical experiments.

Binary Classification from Positive-Confidence Data

Takashi Ishida, Gang Niu, Masashi Sugiyama

Can we learn a binary classifier from only positive data, without any negative data or unlabeled data? We show that if one can equip positive data with confidence (positive-confidence), one can successfully learn a binary classifier, which we name positive-confidence (Pconf) classification. Our work is related to one-class classification which is aimed at "describing" the positive class by clustering-related methods, but one-class classification does not have the ability to tune hyper-parameters and their aim is not on "discriminating" positive and negative classes. For the Pconf classification problem, we provide a simple empirical risk minimization framework that is model-independent and optimization-independent. We theoretically establish the consistency and an estimation error bound, and demonstrate the usefulness of the proposed method for training deep neural networks through experiments.

Transfer Learning with Neural AutoML

Catherine Wong, Neil Houlsby, Yifeng Lu, Andrea Gesmundo

We reduce the computational cost of Neural AutoML with transfer learning. AutoML relieves human effort by automating the design of ML algorithms. Neural AutoML has become popular for the design of deep learning architectures, however, this method has a high computation cost. To address this we propose Transfer Neural AutoML that uses knowledge from prior tasks to speed up network design. We extend RL-based architecture search methods to support parallel training on multiple tasks and then transfer the search strategy to new tasks. On language and image classification data, Transfer Neural AutoML reduces convergence time over single-task training by over an order of magnitude on many tasks.

.

Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikin, Dmitry P. Vetrov, Andrew G. 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.

Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making

Hoda Heidari, Claudio Ferrari, Krishna Gummadi, Andreas Krause

We draw attention to an important, yet largely overlooked aspect of evaluating fairness for automated decision making systems---namely risk and welfare considerations. Our proposed family of measures corresponds to the long-established formulations of cardinal social welfare in economics, and is justified by the Rawlsian conception of fairness behind a veil of ignorance. The convex formulation of our welfare-based measures of fairness allows us to integrate them as a constraint into any convex loss minimization pipeline. Our empirical analysis reveals interesting trade-offs between our proposal and (a) prediction accuracy, (b) group discrimination, and (c) Dwork et al's notion of individual fairness. Furthermore and perhaps most importantly, our work provides both heuristic justification and empirical evidence suggesting that a lower-bound on our measures often leads to bounded inequality in algorithmic outcomes; hence presenting the first computationally feasible mechanism for bounding individual-level inequality.

A Unified View of Piecewise Linear Neural Network Verification

Rudy R. Bunel, Ilker Turkaslan, Philip Torr, Pushmeet Kohli, Pawan K. Mudigonda

The success of Deep Learning and its potential use in many safety-critical applications has motivated research on formal verification of Neural Network (NN) models. Despite the reputation of learned NN models to behave as black boxes and the theoretical hardness of proving their properties, researchers have been successful in verifying some classes of models by exploiting their piecewise linear structure and taking insights from formal methods such as Satisfiability Modulo Theory. These methods are however still far from scaling to realistic neural networks. To facilitate progress on this crucial area, we make two key contributions. First, we present a unified framework that encompasses previous methods. This analysis results in the identification of new methods that combine the strengths of multiple existing approaches, accomplishing a speedup of two orders of magnitude compared to the previous state of the art. Second, we propose a new data set of benchmarks which includes a collection of previously released testcases. We use the benchmark to provide the first experimental comparison of existing algorithms and identify the factors impacting the hardness of verification problems.

Lifted Weighted Mini-Bucket

Nicholas Gallo, Alexander T. Ihler

Many graphical models, such as Markov Logic Networks (MLNs) with evidence, possess highly symmetric substructures but no exact symmetries. Unfortunately, there are few principled methods that exploit these symmetric substructures to perform efficient approximate inference. In this paper, we present a lifted variant of the Weighted Mini-Bucket elimination algorithm which provides a principled way to (i) exploit the highly symmetric substructure of MLN models, and (ii) incorporate high-order inference terms which are necessary for high quality approximate inference. Our method has significant control over the accuracy-time trade-off of the approximation, allowing us to generate any-time approximations. Experimental results demonstrate the utility of this class of approximations, especially in models with strong repulsive potentials.

Can We Gain More from Orthogonality Regularizations in Training Deep Networks?

Nitin Bansal, Xiaohan Chen, Zhangyang Wang

This paper seeks to answer the question: as the (near-) orthogonality of weights is found to be a favorable property for training deep convolutional neural networks, how can we enforce it in more effective and easy-to-use ways? We develop novel orthogonality regularizations on training deep CNNs, utilizing various advanced analytical tools such as mutual coherence and restricted isometry property. These plug-and-play regularizations can be conveniently incorporated into training almost any CNN without extra hassle. We then benchmark their effects on stat

state-of-the-art models: ResNet, WideResNet, and ResNeXt, on several most popular computer vision datasets: CIFAR-10, CIFAR-100, SVHN and ImageNet. We observe consistent performance gains after applying those proposed regularizations, in terms of both the final accuracies achieved, and faster and more stable convergences. We have made our codes and pre-trained models publicly available: <https://github.com/nbansal90/Can-we-Gain-More-from-Orthogonality>.

Understanding the Role of Adaptivity in Machine Teaching: The Case of Version Space Learners

Yuxin Chen, Adish Singla, Oisín Mac Aodha, Pietro Perona, Yisong Yue

In real-world applications of education, an effective teacher adaptively chooses the next example to teach based on the learner's current state. However, most existing work in algorithmic machine teaching focuses on the batch setting, where adaptivity plays no role. In this paper, we study the case of teaching consistent, version space learners in an interactive setting. At any time step, the teacher provides an example, the learner performs an update, and the teacher observes the learner's new state. We highlight that adaptivity does not speed up the teaching process when considering existing models of version space learners, such as the "worst-case" model (the learner picks the next hypothesis randomly from the version space) and the "preference-based" model (the learner picks hypothesis according to some global preference). Inspired by human teaching, we propose a new model where the learner picks hypotheses according to some local preference defined by the current hypothesis. We show that our model exhibits several desirable properties, e.g., adaptivity plays a key role, and the learner's transitions over hypotheses are smooth/interpretable. We develop adaptive teaching algorithms, and demonstrate our results via simulation and user studies.

Wavelet regression and additive models for irregularly spaced data

Asad Haris, Ali Shojaie, Noah Simon

We present a novel approach for nonparametric regression using wavelet basis functions. Our proposal, waveMesh, can be applied to non-equispaced data with sample size not necessarily a power of 2. We develop an efficient proximal gradient descent algorithm for computing the estimator and establish adaptive minimax convergence rates. The main appeal of our approach is that it naturally extends to additive and sparse additive models for a potentially large number of covariates.

We prove minimax optimal convergence rates under a weak compatibility condition for sparse additive models. The compatibility condition holds when we have a small number of covariates. Additionally, we establish convergence rates for when the condition is not met. We complement our theoretical results with empirical studies comparing waveMesh to existing methods.

Self-Erasing Network for Integral Object Attention

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.

Training deep learning based denoisers without ground truth data

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 minimize the mean squared error (MSE) between the output image of a deep neural network and 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.

Structural Causal Bandits: Where to Intervene?

Sanghack Lee, Elias Bareinboim

We study the problem of identifying the best action in a sequential decision-making setting when the reward distributions of the arms exhibit a non-trivial dependence structure, which is governed by the underlying causal model of the domain where the agent is deployed. In this setting, playing an arm corresponds to intervening on a set of variables and setting them to specific values. In this paper, we show that whenever the underlying causal model is not taken into account during the decision-making process, the standard strategies of simultaneously intervening on all variables or on all the subsets of the variables may, in general, lead to suboptimal policies, regardless of the number of interventions performed by the agent in the environment. We formally acknowledge this phenomenon and investigate structural properties implied by the underlying causal model, which lead to a complete characterization of the relationships between the arms' distributions. We leverage this characterization to build a new algorithm that takes as input a causal structure and finds a minimal, sound, and complete set of qualified arms that an agent should play to maximize its expected reward. We empirically demonstrate that the new strategy learns an optimal policy and leads to orders of magnitude faster convergence rates when compared with its causal-insensitive counterparts.

Scalar Posterior Sampling with Applications

Georgios Theodorou, Zheng Wen, Yasin Abbasi-Yadkori, Nikos Vlassis

We propose a practical non-episodic PSRL algorithm that unlike recent state-of-the-art PSRL algorithms uses a deterministic, model-independent episode switching schedule. Our algorithm termed deterministic schedule PSRL (DS-PSRL) is efficient in terms of time, sample, and space complexity. We prove a Bayesian regret bound under mild assumptions. Our result is more generally applicable to multiple parameters and continuous state action problems. We compare our algorithm with state-of-the-art PSRL algorithms on standard discrete and continuous problems from the literature. Finally, we show how the assumptions of our algorithm satisfy a sensible parameterization for a large class of problems in sequential recommendations.

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

he 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.

Online Learning of Quantum States

Scott Aaronson, Xinyi Chen, Elad Hazan, Satyen Kale, Ashwin Nayak

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

Avital Oliver, Augustus Odena, Colin A. Raffel, Ekin Dogus Cubuk, Ian Goodfellow
Semi-supervised learning (SSL) provides a powerful framework for leveraging unlabeled data when labels are limited or expensive to obtain. SSL algorithms based on deep neural networks have recently proven successful on standard benchmark tasks. However, we argue that these benchmarks fail to address many issues that SSL algorithms would face in real-world applications. After creating a unified reimplementation of various widely-used SSL techniques, we test them in a suite of experiments designed to address these issues. We find that the performance of simple baselines which do not use unlabeled data is often underreported, SSL methods differ in sensitivity to the amount of labeled and unlabeled data, and performance can degrade substantially when the unlabeled dataset contains out-of-distribution examples. To help guide SSL research towards real-world applicability, we make our unified reimplementation and evaluation platform publicly available.

Long short-term memory and Learning-to-learn in networks of spiking neurons

Guillaume Bellec, Darjan Salaj, Anand Subramoney, Robert Legenstein, Wolfgang Maass

Recurrent networks of spiking neurons (RSNNs) underlie the astounding computing and learning capabilities of the brain. But computing and learning capabilities of RSNN models have remained poor, at least in comparison with ANNs. We address two possible reasons for that. One is that RSNNs in the brain are not randomly connected or designed according to simple rules, and they do not start learning as a tabula rasa network. Rather, RSNNs in the brain were optimized for their tasks through evolution, development, and prior experience. Details of these optimization processes are largely unknown. But their functional contribution can be approximated through powerful optimization methods, such as backpropagation through time (BPTT).

Revisiting Decomposable Submodular Function Minimization with Incidence Relations

Pan Li, Olga Milenkovic

We introduce a new approach to decomposable submodular function minimization (DSFM) that exploits incidence relations. Incidence relations describe which variables effectively influence the component functions, and when properly utilized, they allow for improving the convergence rates of DSFM solvers. Our main results include the precise parametrization of the DSFM problem based on incidence relations, the development of new scalable alternative projections and parallel coordinate descent methods and an accompanying rigorous analysis of their convergence

rates.

DifNet: Semantic Segmentation by Diffusion Networks

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. Our DifNet is trained end-to-end without complex loss functions.

Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering

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 to reason 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.

Distributed Multi-Player Bandits - a Game of Thrones Approach

Ilai Bistriz, Amir Leshem

We consider a multi-armed bandit game where N players compete for K arms for T turns. Each player has different expected rewards for the arms, and the instantaneous rewards are independent and identically distributed. Performance is measured using the expected sum of regrets, compared to the optimal assignment of arms to players. We assume that each player only knows her actions and the reward she received each turn. Players cannot observe the actions of other players, and no communication between players is possible. We present a distributed algorithm and prove that it achieves an expected sum of regrets of $\text{near-}O(\log^2 T)$. This is the first algorithm to achieve a poly-logarithmic regret in this fully distributed scenario. All other works have assumed that either all players have the same vector of expected rewards or that communication between players is possible.

Scaling Gaussian Process Regression with Derivatives

David Eriksson, Kun Dong, Eric Lee, David Bindel, Andrew G. Wilson

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors.

ors prior to requesting a name change in the electronic proceedings.

Deep Attentive Tracking via Reciprocatative Learning

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 reciprocal 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.

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

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.

Global Geometry of Multichannel Sparse Blind Deconvolution on the Sphere

YanJun Li, Yoram Bresler

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Sample Efficient Stochastic Gradient Iterative Hard Thresholding Method for Stochastic Sparse Linear Regression with Limited Attribute Observation

Tomoya Murata, Taiji Suzuki

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Blockwise Parallel Decoding for Deep Autoregressive Models

Mitchell Stern, Noam Shazeer, Jakob Uszkoreit

Deep autoregressive sequence-to-sequence models have demonstrated impressive performance across a wide variety of tasks in recent years. While common architecture classes such as recurrent, convolutional, and self-attention networks make different trade-offs between the amount of computation needed per layer and the le

length of the critical path at training time, generation still remains an inherently sequential process. To overcome this limitation, we propose a novel blockwise parallel decoding scheme in which we make predictions for multiple time steps in parallel then back off to the longest prefix validated by a scoring model. This allows for substantial theoretical improvements in generation speed when applied to architectures that can process output sequences in parallel. We verify our approach empirically through a series of experiments using state-of-the-art self-attention models for machine translation and image super-resolution, achieving iteration reductions of up to 2x over a baseline greedy decoder with no loss in quality, or up to 7x in exchange for a slight decrease in performance. In terms of wall-clock time, our fastest models exhibit real-time speedups of up to 4x over standard greedy decoding.

Sublinear Time Low-Rank Approximation of Distance Matrices

Ainesh Bakshi, David Woodruff

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization

Bruno Korbar, Du Tran, Lorenzo Torresani

There is a natural correlation between the visual and auditive elements of a video. In this work we leverage this connection to learn general and effective models for both audio and video analysis from self-supervised temporal synchronization. We demonstrate that a calibrated curriculum learning scheme, a careful choice of negative examples, and the use of a contrastive loss are critical ingredients to obtain powerful multi-sensory representations from models optimized to discern temporal synchronization of audio-video pairs. Without further fine-tuning, the resulting audio features achieve performance superior or comparable to the state-of-the-art on established audio classification benchmarks (DCASE2014 and ESC-50). At the same time, our visual subnet provides a very effective initialization to improve the accuracy of video-based action recognition models: compared to learning from scratch, our self-supervised pretraining yields a remarkable gain of +19.9% in action recognition accuracy on UCF101 and a boost of +17.7% on HMDB51.

Submodular Field Grammars: Representation, Inference, and Application to Image Parsing

Abram L. Friesen, Pedro M. Domingos

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification

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 image

s 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.

Estimators for Multivariate Information Measures in General Probability Spaces

Arman Rahimzamani, Himanshu Asnani, Pramod Viswanath, Sreeram Kannan

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Graphical Generative Adversarial Networks

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 Homogeneous Mixture Models: Representation, Separation, and Approximation

Priyank Jaini, Pascal Poupart, Yaoliang Yu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives

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 *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 *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.

Community Exploration: From Offline Optimization to Online Learning

Xiaowei Chen, Weiran Huang, Wei Chen, John C. S. Lui

We introduce the community exploration problem that has various real-world applications such as online advertising. In the problem, an explorer allocates limited budget to explore communities so as to maximize the number of members he could meet. We provide a systematic study of the community exploration problem, from offline optimization to online learning. For the offline setting where the sizes

of communities are known, we prove that the greedy methods for both of non-adaptive exploration and adaptive exploration are optimal. For the online setting where the sizes of communities are not known and need to be learned from the multi-round explorations, we propose an ``upper confidence'' like algorithm that achieves the logarithmic regret bounds. By combining the feedback from different rounds, we can achieve a constant regret bound.

Context-aware Synthesis and Placement of Object Instances

Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, 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.

Beyond Log-concavity: Provable Guarantees for Sampling Multi-modal Distributions using Simulated Tempering Langevin Monte Carlo

Holden Lee, Andrej Risteski, Rong Ge

A key task in Bayesian machine learning is sampling from distributions that are only specified up to a partition function (i.e., constant of proportionality). One prevalent example of this is sampling posteriors in parametric distributions, such as latent-variable generative models. However sampling (even very approximately) can be $\#P$ -hard.

M-Walk: Learning to Walk over Graphs using Monte Carlo Tree Search

Yelong Shen, Jianshu Chen, Po-Sen Huang, Yuqing Guo, Jianfeng Gao

Learning to walk over a graph towards a target node for a given query and a source node is an important problem in applications such as knowledge base completion (KBC). It can be formulated as a reinforcement learning (RL) problem with a known state transition model. To overcome the challenge of sparse rewards, we develop a graph-walking agent called M-Walk, which consists of a deep recurrent neural network (RNN) and Monte Carlo Tree Search (MCTS). The RNN encodes the state (i.e., history of the walked path) and maps it separately to a policy and Q-values. In order to effectively train the agent from sparse rewards, we combine MCTS with the neural policy to generate trajectories yielding more positive rewards. From these trajectories, the network is improved in an off-policy manner using Q-learning, which modifies the RNN policy via parameter sharing. Our proposed RL algorithm repeatedly applies this policy-improvement step to learn the model. At test time, MCTS is combined with the neural policy to predict the target node. Experimental results on several graph-walking benchmarks show that M-Walk is able to learn better policies than other RL-based methods, which are mainly based on policy gradients. M-Walk also outperforms traditional KBC baselines.

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

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.

Group Equivariant Capsule Networks

Jan Eric Lenssen, Matthias Fey, Pascal Libuschewski

We present group equivariant capsule networks, a framework to introduce guaranteed equivariance and invariance properties to the capsule network idea. Our work can be divided into two contributions. First, we present a generic routing by agreement algorithm defined on elements of a group and prove that equivariance of output pose vectors, as well as invariance of output activations, hold under certain conditions. Second, we connect the resulting equivariant capsule networks with work from the field of group convolutional networks. Through this connection, we provide intuitions of how both methods relate and are able to combine the strengths of both approaches in one deep neural network architecture. The resulting framework allows sparse evaluation of the group convolution operator, provides control over specific equivariance and invariance properties, and can use routing by agreement instead of pooling operations. In addition, it is able to provide interpretable and equivariant representation vectors as output capsules, which disentangle evidence of object existence from its pose.

Fairness Through Computationally-Bounded Awareness

Michael Kim, Omer Reingold, Guy Rothblum

We study the problem of fair classification within the versatile framework of Dwork et al. [ITCS '12], which assumes the existence of a metric that measures similarity between pairs of individuals. Unlike earlier work, we do not assume that the entire metric is known to the learning algorithm; instead, the learner can query this arbitrary metric a bounded number of times. We propose a new notion of fairness called metric multifairness and show how to achieve this notion in our setting.

Metric multifairness is parameterized by a similarity metric d on pairs of individuals to classify and a rich collection C of (possibly overlapping) "comparison sets" over pairs of individuals. At a high level, metric multifairness guarantees that similar subpopulations are treated similarly, as long as these subpopulations are identified within the class C .

Geometry Based Data Generation

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 even 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.

Searching for Efficient Multi-Scale Architectures for Dense Image Prediction

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.

Adversarial Scene Editing: Automatic Object Removal from Weak Supervision

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.

Variational Inverse Control with Events: A General Framework for Data-Driven Reward Definition

Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, Sergey Levine

The design of a reward function often poses a major practical challenge to real-world applications of reinforcement learning. Approaches such as inverse reinforcement learning attempt to overcome this challenge, but require expert demonstrations, which can be difficult or expensive to obtain in practice. We propose inverse event-based control, which generalizes inverse reinforcement learning methods to cases where full demonstrations are not needed, such as when only samples of desired goal states are available. Our method is grounded in an alternative perspective on control and reinforcement learning, where an agent's goal is to maximize the probability that one or more events will happen at some point in the future, rather than maximizing cumulative rewards. We demonstrate the effectiveness of our methods on continuous control tasks, with a focus on high-dimensional observations like images where rewards are hard or even impossible to specify.

MULAN: A Blind and Off-Grid Method for Multichannel Echo Retrieval

Helena Peic Tukuljac, Antoine Deleforge, Remi Gribonval

This paper addresses the general problem of blind echo retrieval, i.e., given M sensors measuring in the discrete-time domain M mixtures of K delayed and attenuated

ated copies of an unknown source signal, can the echo location and weights be recovered? This problem has broad applications in fields such as sonars, seismology, ultrasounds or room acoustics. It belongs to the broader class of blind channel identification problems, which have been intensively studied in signal processing. All existing methods proceed in two steps: (i) blind estimation of sparse discrete-time filters and (ii) echo information retrieval by peak picking. The precision of these methods is fundamentally limited by the rate at which the signals are sampled: estimated echo locations are necessary on-grid, and since true locations never match the sampling grid, the weight estimation precision is also strongly limited. This is the so-called basis-mismatch problem in compressed sensing. We propose a radically different approach to the problem, building on top of the framework of finite-rate-of-innovation sampling. The approach operates directly in the parameter-space of echo locations and weights, and enables near-exact blind and off-grid echo retrieval from discrete-time measurements. It is shown to outperform conventional methods by several orders of magnitudes in precision.

Diminishing Returns Shape Constraints for Interpretability and Regularization

Maya Gupta, Dara Bahri, Andrew Cotter, Kevin Canini

We investigate machine learning models that can provide diminishing returns and accelerating returns guarantees to capture prior knowledge or policies about how outputs should depend on inputs. We show that one can build flexible, nonlinear, multi-dimensional models using lattice functions with any combination of concavity/convexity and monotonicity constraints on any subsets of features, and compare to new shape-constrained neural networks. We demonstrate on real-world examples that these shape constrained models can provide tuning-free regularization and improve model understandability.

Breaking the Activation Function Bottleneck through Adaptive Parameterization

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.

Sketching Method for Large Scale Combinatorial Inference

Wei Sun, Junwei Lu, Han Liu

We present computationally efficient algorithms to test various combinatorial structures of large-scale graphical models. In order to test the hypotheses on their topological structures, we propose two adjacency matrix sketching frameworks: neighborhood sketching and subgraph sketching. The neighborhood sketching algorithm is proposed to test the connectivity of graphical models. This algorithm randomly subsamples vertices and conducts neighborhood regression and screening. The global sketching algorithm is proposed to test the topological properties requiring exponential computation complexity, especially testing the chromatic number and the maximum clique. This algorithm infers the corresponding property based on the sampled subgraph. Our algorithms are shown to substantially accelerate the computation of existing methods. We validate our theory and method through both synthetic simulations and a real application in neuroscience.

Symbolic Graph Reasoning Meets Convolutions

Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, Eric P. Xing

Beyond local convolution networks, we explore how to harness various external human knowledge for endowing the networks with the capability of semantic global reasoning. Rather than using separate graphical models (e.g. CRF) or constraints for modeling broader dependencies, we propose a new Symbolic Graph Reasoning (SG

R) layer, which performs reasoning over a group of symbolic nodes whose outputs explicitly represent different properties of each semantic in a prior knowledge graph. To cooperate with local convolutions, each SGR is constituted by three modules: a) a primal local-to-semantic voting module where the features of all symbolic nodes are generated by voting from local representations; b) a graph reasoning module propagates information over knowledge graph to achieve global semantic coherency; c) a dual semantic-to-local mapping module learns new associations of the evolved symbolic nodes with local representations, and accordingly enhances local features. The SGR layer can be injected between any convolution layers and instantiated with distinct prior graphs. Extensive experiments show incorporating SGR significantly improves plain ConvNets on three semantic segmentation tasks and one image classification task. More analyses show the SGR layer learns shared symbolic representations for domains/datasets with the different label set given a universal knowledge graph, demonstrating its superior generalization capability.

Topkapi: Parallel and Fast Sketches for Finding Top-K Frequent Elements
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.

The Price of Fair PCA: One Extra dimension

Samira Samadi, Uthaipon Tantipongpipat, Jamie H. 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.

Orthogonally Decoupled Variational Gaussian Processes

Hugh Salimbeni, Ching-An Cheng, Byron Boots, Marc Deisenroth

Gaussian processes (GPs) provide a powerful non-parametric framework for reasoning over functions. Despite appealing theory, its superlinear computational and memory complexities have presented a long-standing challenge. State-of-the-art sparse variational inference methods trade modeling accuracy against complexity. However, the complexities of these methods still scale superlinearly in the number of basis functions, implying that that sparse GP methods are able to learn from large datasets only when a small model is used. Recently, a decoupled approach was proposed that removes the unnecessary coupling between the complexities of modeling the mean and the covariance functions of a GP. It achieves a linear complexity in the number of mean parameters, so an expressive posterior mean function

ion can be modeled. While promising, this approach suffers from optimization difficulties due to ill-conditioning and non-convexity. In this work, we propose an alternative decoupled parametrization. It adopts an orthogonal basis in the mean function to model the residues that cannot be learned by the standard coupled approach. Therefore, our method extends, rather than replaces, the coupled approach to achieve strictly better performance. This construction admits a straightforward natural gradient update rule, so the structure of the information manifold that is lost during decoupling can be leveraged to speed up learning. Empirically, our algorithm demonstrates significantly faster convergence in multiple experiments.

Algorithmic Assurance: An Active Approach to Algorithmic Testing using Bayesian Optimisation

Shivapratap Gopakumar, Sunil Gupta, Santu Rana, Vu Nguyen, Svetha Venkatesh

We introduce algorithmic assurance, the problem of testing whether machine learning algorithms are conforming to their intended design goal. We address this problem by proposing an efficient framework for algorithmic testing. To provide assurance, we need to efficiently discover scenarios where an algorithm decision deviates maximally from its intended gold standard. We mathematically formulate this task as an optimisation problem of an expensive, black-box function. We use an active learning approach based on Bayesian optimisation to solve this optimisation problem. We extend this framework to algorithms with vector-valued outputs by making appropriate modification in Bayesian optimisation via the EXP3 algorithm. We theoretically analyse our methods for convergence. Using two real-world applications, we demonstrate the efficiency of our methods. The significance of our problem formulation and initial solutions is that it will serve as the foundation in assuring humans about machines making complex decisions.

Domain-Invariant Projection Learning for Zero-Shot Recognition

An Zhao, Mingyu Ding, Jiechao Guan, Zhiwu Lu, Tao Xiang, Ji-Rong Wen

Zero-shot learning (ZSL) aims to recognize unseen object classes without any training samples, which can be regarded as a form of transfer learning from seen classes to unseen ones. This is made possible by learning a projection between a feature space and a semantic space (e.g. attribute space). Key to ZSL is thus to learn a projection function that is robust against the often large domain gap between the seen and unseen classes. In this paper, we propose a novel ZSL model termed domain-invariant projection learning (DIPL). Our model has two novel components: (1) A domain-invariant feature self-reconstruction task is introduced to the seen/unseen class data, resulting in a simple linear formulation that casts ZSL into a min-min optimization problem. Solving the problem is non-trivial, and a novel iterative algorithm is formulated as the solver, with rigorous theoretical algorithm analysis provided. (2) To further align the two domains via the learned projection, shared semantic structure among seen and unseen classes is explored via forming superclasses in the semantic space. Extensive experiments show that our model outperforms the state-of-the-art alternatives by significant margins.

High Dimensional Linear Regression using Lattice Basis Reduction

Ilias Zadik, David Gamarnik

We consider a high dimensional linear regression problem where the goal is to efficiently recover an unknown vector β^* from n noisy linear observations $Y = X\beta^* + W$ in \mathbb{R}^n , for known X in $\mathbb{R}^{\{n\} \times p}$ and unknown W in \mathbb{R}^n . Unlike most of the literature on this model we make no sparsity assumption on β^* . Instead we adopt a regularization based on assuming that the underlying vectors β^* have rational entries with the same denominator Q . We call this Q -rationality assumption. We propose a new polynomial-time algorithm for this task which is based on the seminal Lenstra-Lenstra-Lovasz (LLL) lattice basis reduction algorithm. We establish that under the Q -rationality assumption, our algorithm reco

vers exactly the vector β^* for a large class of distributions for the iid entries of X and non-zero noise W . We prove that it is successful under small noise, even when the learner has access to only one observation ($n=1$). Furthermore, we prove that in the case of the Gaussian white noise for W , $n=o(p/\log p)$ and Q sufficiently large, our algorithm tolerates a nearly optimal information-theoretic level of the noise.

A Retrieve-and-Edit Framework for Predicting Structured Outputs

Tatsunori B. Hashimoto, Kelvin Guu, Yonatan Oren, Percy S. 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.

Efficient inference for time-varying behavior during learning

Nicholas A. Roy, Ji Hyun Bak, Athena Akrami, Carlos Brody, Jonathan W. Pillow

The process of learning new behaviors over time is a problem of great interest in both neuroscience and artificial intelligence. However, most standard analyses of animal training data either treat behavior as fixed or track only coarse performance statistics (e.g., accuracy, bias), providing limited insight into the evolution of the policies governing behavior. To overcome these limitations, we propose a dynamic psychophysical model that efficiently tracks trial-to-trial changes in behavior over the course of training. Our model consists of a dynamic logistic regression model, parametrized by a set of time-varying weights that express dependence on sensory stimuli as well as task-irrelevant covariates, such as stimulus, choice, and answer history. Our implementation scales to large behavioral datasets, allowing us to infer 500K parameters (e.g. 10 weights over 50K trials) in minutes on a desktop computer. We optimize hyperparameters governing how rapidly each weight evolves over time using the decoupled Laplace approximation, an efficient method for maximizing marginal likelihood in non-conjugate models. To illustrate performance, we apply our method to psychophysical data from both rats and human subjects learning a delayed sensory discrimination task. The model successfully tracks the psychophysical weights of rats over the course of training, capturing day-to-day and trial-to-trial fluctuations that underlie changes in performance, choice bias, and dependencies on task history. Finally, we investigate why rats frequently make mistakes on easy trials, and suggest that apparent lapses can be explained by sub-optimal weighting of known task covariates.

Re-evaluating evaluation

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 ea

sy tasks or weak agents. Nash averaging thus encourages maximally inclusive evaluation -- since there is no harm (computational cost aside) from including all a vailable tasks and agents.

Learning Abstract Options

Matthew Riemer, Miao Liu, Gerald Tesauro

Building systems that autonomously create temporal abstractions from data is a key challenge in scaling learning and planning in reinforcement learning. One popular approach for addressing this challenge is the options framework (Sutton et al., 1999). However, only recently in (Bacon et al., 2017) was a policy gradient theorem derived for online learning of general purpose options in an end to end fashion. In this work, we extend previous work on this topic that only focuses on learning a two-level hierarchy including options and primitive actions to enable learning simultaneously at multiple resolutions in time. We achieve this by considering an arbitrarily deep hierarchy of options where high level temporally extended options are composed of lower level options with finer resolutions in time. We extend results from (Bacon et al., 2017) and derive policy gradient theorems for a deep hierarchy of options. Our proposed hierarchical option-critic architecture is capable of learning internal policies, termination conditions, and hierarchical compositions over options without the need for any intrinsic rewards or subgoals. Our empirical results in both discrete and continuous environments demonstrate the efficiency of our framework.

Optimal Algorithms for Continuous Non-monotone Submodular and DR-Submodular Maximization

Rad Niazadeh, Tim Roughgarden, Joshua Wang

In this paper we study the fundamental problems of maximizing a continuous non monotone submodular function over a hypercube, with and without coordinate-wise concavity. This family of optimization problems has several applications in machine learning, economics, and communication systems. Our main result is the first $1/2$ approximation algorithm for continuous submodular function maximization; this approximation factor of is the best possible for algorithms that use only polynomially many queries. For the special case of DR-submodular maximization, we provide a faster $1/2$ -approximation algorithm that runs in (almost) linear time. Both of these results improve upon prior work [Bian et al., 2017, Soma and Yoshida, 2017, Buchbinder et al., 2012].

Sequential Context Encoding for Duplicate Removal

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.

PG-TS: Improved Thompson Sampling for Logistic Contextual Bandits

Bianca Dumitrascu, Karen Feng, Barbara Engelhardt

We address the problem of regret minimization in logistic contextual bandits, where a learner decides among sequential actions or arms given their respective contexts to maximize binary rewards. Using a fast inference procedure with Polya-Gamma distributed augmentation variables, we propose an improved version of Thompson Sampling, a Bayesian formulation of contextual bandits with near-optimal per

formance. Our approach, Polya-Gamma augmented Thompson Sampling (PG-TS), achieves state-of-the-art performance on simulated and real data. PG-TS explores the action space efficiently and exploits high-reward arms, quickly converging to solutions of low regret. Its explicit estimation of the posterior distribution of the context feature covariance leads to substantial empirical gains over approximate approaches. PG-TS is the first approach to demonstrate the benefits of Polya-Gamma augmentation in bandits and to propose an efficient Gibbs sampler for approximating the analytically unsolvable integral of logistic contextual bandits.

Variance-Reduced Stochastic Gradient Descent on Streaming Data

Ellango Jothimurugesan, Ashraf Tahmasbi, Phillip Gibbons, Srikanta Tirthapura

We present an algorithm STRSAGA for efficiently maintaining a machine learning model over data points that arrive over time, quickly updating the model as new training data is observed. We present a competitive analysis comparing the suboptimality of the model maintained by STRSAGA with that of an offline algorithm that is given the entire data beforehand, and analyze the risk-competitiveness of STRSAGA under different arrival patterns. Our theoretical and experimental results show that the risk of STRSAGA is comparable to that of offline algorithms on a variety of input arrival patterns, and its experimental performance is significantly better than prior algorithms suited for streaming data, such as SGD and SVRG.

Reinforced Continual Learning

Ju Xu, Zhanxing Zhu

Most artificial intelligence models are limited in their ability to solve new tasks faster, without forgetting previously acquired knowledge. The recently emerging paradigm of continual learning aims to solve this issue, in which the model learns various tasks in a sequential fashion. In this work, a novel approach for continual learning is proposed, which searches for the best neural architecture for each coming task via sophisticatedly designed reinforcement learning strategies. We name it as Reinforced Continual Learning. Our method not only has good performance on preventing catastrophic forgetting but also fits new tasks well. The experiments on sequential classification tasks for variants of MNIST and CIFAR-100 datasets demonstrate that the proposed approach outperforms existing continual learning alternatives for deep networks.

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

Mingchao Yu, Zhifeng Lin, Krishna Narra, Songze Li, Youjie Li, Nam Sung Kim, Alexander Schwing, Murali Annamalai, 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.

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

resholds

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>.

Optimization for Approximate Submodularity

Yaron Singer, Avinatan Hassidim

We consider the problem of maximizing a submodular function when given access to its approximate version. Submodular functions are heavily studied in a wide variety of disciplines, since they are used to model many real world phenomena, and are amenable to optimization. However, there are many cases in which the phenomena we observe is only approximately submodular and the approximation guarantees cease to hold. We describe a technique which we call the sampled mean approximation that yields strong guarantees for maximization of submodular functions from approximate surrogates under cardinality and intersection of matroid constraints. In particular, we show tight guarantees for maximization under a cardinality constraint and $1/(1+P)$ approximation under intersection of P matroids.

Efficient Neural Network Robustness Certification with General Activation Functions

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, Luca Daniel

Finding minimum distortion of adversarial examples and thus certifying robustness in neural networks classifiers is known to be a challenging problem. Nevertheless, recently it has been shown to be possible to give a non-trivial certified lower bound of minimum distortion, and some recent progress has been made towards this direction by exploiting the piece-wise linear nature of ReLU activations. However, a generic robustness certification for general activation functions still remains largely unexplored. To address this issue, in this paper we introduce CROWN, a general framework to certify robustness of neural networks with general activation functions. The novelty in our algorithm consists of bounding a given activation function with linear and quadratic functions, hence allowing it to tackle general activation functions including but not limited to the four popular choices: ReLU, tanh, sigmoid and arctan. In addition, we facilitate the search for a tighter certified lower bound by adaptively selecting appropriate surrogates for each neuron activation. Experimental results show that CROWN on ReLU networks can notably improve the certified lower bounds compared to the current state-of-the-art algorithm Fast-Lin, while having comparable computational efficiency. Furthermore, CROWN also demonstrates its effectiveness and flexibility on networks with general activation functions, including tanh, sigmoid and arctan.

Neural Edit Operations for Biological Sequences

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

st 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.

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

Tyler Scott, Karl Ridgeway, Michael C. 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 finetune 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.

The Importance of Sampling in Meta-Reinforcement Learning

Bradly Stadie, Ge Yang, Rein Houthoofd, Peter Chen, Yan Duan, Yuhuai Wu, Pieter Abbeel, Ilya Sutskever

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

KONG: Kernels for ordered-neighborhood graphs

Moez Draief, Konstantin Kutikov, Kevin Scaman, Milan Vojnovic

We present novel graph kernels for graphs with node and edge labels that have ordered neighborhoods, i.e. when neighbor nodes follow an order. Graphs with ordered neighborhoods are a natural data representation for evolving graphs where edges are created over time, which induces an order. Combining convolutional subgraph kernels and string kernels, we design new scalable algorithms for generation of explicit graph feature maps using sketching techniques. We obtain precise bounds for the approximation accuracy and computational complexity of the proposed approaches and demonstrate their applicability on real datasets. In particular, our experiments demonstrate that neighborhood ordering results in more informative features. For the special case of general graphs, i.e. graphs without ordered neighborhoods, the new graph kernels yield efficient and simple algorithms for the comparison of label distributions between graphs.

Glow: Generative Flow with Invertible 1x1 Convolutions

Durk P. 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.

Efficient Projection onto the Perfect Phylogeny Model

Bei Jia, Surjyendu Ray, Sam Safavi, José Bento

Several algorithms build on the perfect phylogeny model to infer evolutionary trees. This problem is particularly hard when evolutionary trees are inferred from the fraction of genomes that have mutations in different positions, across different samples. Existing algorithms might do extensive searches over the space of possible trees. At the center of these algorithms is a projection problem that assigns a fitness cost to phylogenetic trees. In order to perform a wide search over the space of the trees, it is critical to solve this projection problem fast. In this paper, we use Moreau's decomposition for proximal operators, and a tree reduction scheme, to develop a new algorithm to compute this projection. Our algorithm terminates with an exact solution in a finite number of steps, and is extremely fast. In particular, it can search over all evolutionary trees with fewer than 11 nodes, a size relevant for several biological problems (more than 2 billion trees) in about 2 hours.

Do Less, Get More: Streaming Submodular Maximization with Subsampling

Moran Feldman, Amin Karbasi, Ehsan Kazemi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Temporal alignment and latent Gaussian process factor inference in population spike trains

Lea Duncker, Maneesh Sahani

We introduce a novel scalable approach to identifying common latent structure in neural population spike-trains, which allows for variability both in the trajectory and in the rate of progression of the underlying computation. Our approach is based on shared latent Gaussian processes (GPs) which are combined linearly, as in the Gaussian Process Factor Analysis (GPFA) algorithm. We extend GPFA to handle unbinned spike-train data by incorporating a continuous time point-process likelihood model, achieving scalability with a sparse variational approximation. Shared variability is separated into terms that express condition dependence, as well as trial-to-trial variation in trajectories. Finally, we introduce a nested GP formulation to capture variability in the rate of evolution along the trajectory. We show that the new method learns to recover latent trajectories in synthetic data, and can accurately identify the trial-to-trial timing of movement-related parameters from motor cortical data without any supervision.

Dimensionality Reduction for Stationary Time Series via Stochastic Nonconvex Optimization

Minshuo Chen, Lin Yang, Mengdi Wang, Tuo Zhao

Stochastic optimization naturally arises in machine learning. Efficient algorithms with provable guarantees, however, are still largely missing, when the objective function is nonconvex and the data points are dependent. This paper studies this fundamental challenge through a streaming PCA problem for stationary time series data. Specifically, our goal is to estimate the principle component of time series data with respect to the covariance matrix of the stationary distribution. Computationally, we propose a variant of Oja's algorithm combined with downs

amplifying to control the bias of the stochastic gradient caused by the data dependency. Theoretically, we quantify the uncertainty of our proposed stochastic algorithm based on diffusion approximations. This allows us to prove the asymptotic rate of convergence and further implies near optimal asymptotic sample complexity. Numerical experiments are provided to support our analysis.

Nonlocal Neural Networks, Nonlocal Diffusion and Nonlocal Modeling

Yunzhe Tao, Qi Sun, Qiang Du, Wei Liu

Nonlocal neural networks have been proposed and shown to be effective in several computer vision tasks, where the nonlocal operations can directly capture long-range dependencies in the feature space. In this paper, we study the nature of diffusion and damping effect of nonlocal networks by doing spectrum analysis on the weight matrices of the well-trained networks, and then propose a new formulation of the nonlocal block. The new block not only learns the nonlocal interactions but also has stable dynamics, thus allowing deeper nonlocal structures. Moreover, we interpret our formulation from the general nonlocal modeling perspective, where we make connections between the proposed nonlocal network and other nonlocal models, such as nonlocal diffusion process and Markov jump process.

Partially-Supervised Image Captioning

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.

SLANG: Fast Structured Covariance Approximations for Bayesian Deep Learning with Natural Gradient

Aaron Mishkin, Frederik Kunstner, Didrik Nielsen, Mark Schmidt, Mohammad Emtiyaz Khan

Uncertainty estimation in large deep-learning models is a computationally challenging

task, where it is difficult to form even a Gaussian approximation to the posterior distribution. In such situations, existing methods usually resort to a diagonal

approximation of the covariance matrix despite the fact that these matrices are known to give poor uncertainty estimates. To address this issue, we propose a new stochastic, low-rank, approximate natural-gradient (SLANG) method for variational inference in large deep models. Our method estimates a "diagonal plus low-rank" structure based solely on back-propagated gradients of the network

log-likelihood. This requires strictly less gradient computations than methods that

compute the gradient of the whole variational objective. Empirical evaluations on standard benchmarks confirm that SLANG enables faster and more accurate estimation of uncertainty than mean-field methods, and performs comparably to

state-of-the-art methods.

On Learning Markov Chains

Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati

The problem of estimating an unknown discrete distribution from its samples is a fundamental tenet of statistical learning. Over the past decade, it attracted significant research effort and has been solved for a variety of divergence measures. Surprisingly, an equally important problem, estimating an unknown Markov chain from its samples, is still far from understood. We consider two problems related to the min-max risk (expected loss) of estimating an unknown k -state Markov chain from its n sequential samples: predicting the conditional distribution of the next sample with respect to the KL-divergence, and estimating the transition matrix with respect to a natural loss induced by KL or a more general f -divergence measure.

Is Q-Learning Provably Efficient?

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, Michael I. Jordan

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Preference Based Adaptation for Learning Objectives

Yao-Xiang Ding, Zhi-Hua Zhou

In many real-world learning tasks, it is hard to directly optimize the true performance measures, meanwhile choosing the right surrogate objectives is also difficult. Under this situation, it is desirable to incorporate an optimization of objective process into the learning loop based on weak modeling of the relationship between the true measure and the objective. In this work, we discuss the task of objective adaptation, in which the learner iteratively adapts the learning objective to the underlying true objective based on the preference feedback from an oracle. We show that when the objective can be linearly parameterized, this preference based learning problem can be solved by utilizing the dueling bandit model. A novel sampling based algorithm DL^2M is proposed to learn the optimal parameter, which enjoys strong theoretical guarantees and efficient empirical performance. To avoid learning a hypothesis from scratch after each objective function update, a boosting based hypothesis adaptation approach is proposed to efficiently adapt any pre-learned element hypothesis to the current objective. We apply the overall approach to multi-label learning, and show that the proposed approach achieves significant performance under various multi-label performance measures.

Learning Gaussian Processes by Minimizing PAC-Bayesian Generalization Bounds

David Reeb, Andreas Doerr, Sebastian Gerwinn, Barbara Rakitsch

Gaussian Processes (GPs) are a generic modelling tool for supervised learning. While they have been successfully applied on large datasets, their use in safety-critical applications is hindered by the lack of good performance guarantees. To this end, we propose a method to learn GPs and their sparse approximations by directly optimizing a PAC-Bayesian bound on their generalization performance, instead of maximizing the marginal likelihood. Besides its theoretical appeal, we find in our evaluation that our learning method is robust and yields significantly better generalization guarantees than other common GP approaches on several regression benchmark datasets.

Learning Invariances using the Marginal Likelihood

Mark van der Wilk, Matthias Bauer, ST John, James Hensman

In many supervised learning tasks, learning what changes do not affect the prediction target is as crucial to generalisation as learning what does. Data augmentation is a common way to enforce a model to exhibit an invariance: training data is modified according to an invariance designed by a human and added to the training set.

aining data. We argue that invariances should be incorporated the model structure, and learned using the marginal likelihood, which can correctly reward the reduced complexity of invariant models. We incorporate invariances in a Gaussian process, due to good marginal likelihood approximations being available for these models. Our main contribution is a derivation for a variational inference scheme for invariant Gaussian processes where the invariance is described by a probability distribution that can be sampled from, much like how data augmentation is implemented in practice

NEON2: Finding Local Minima via First-Order Oracles

Zeyuan Allen-Zhu, Yuanzhi Li

We propose a reduction for non-convex optimization that can (1) turn an stationary-point finding algorithm into an local-minimum finding one, and (2) replace the Hessian-vector product computations with only gradient computations. It works both in the stochastic and the deterministic settings, without hurting the algorithm's performance.

Genetic-Gated Networks for Deep Reinforcement Learning

Simyung Chang, John Yang, Jaeseok Choi, Nojun Kwak

We introduce the Genetic-Gated Networks (G2Ns), simple neural networks that combine a gate vector composed of binary genetic genes in the hidden layer(s) of networks. Our method can take both advantages of gradient-free optimization and gradient-based optimization methods, of which the former is effective for problems with multiple local minima, while the latter can quickly find local minima. In addition, multiple chromosomes can define different models, making it easy to construct multiple models and can be effectively applied to problems that require multiple models. We show that this G2N can be applied to typical reinforcement learning algorithms to achieve a large improvement in sample efficiency and performance.

Lipschitz regularity of deep neural networks: analysis and efficient estimation

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.

Accelerated Stochastic Matrix Inversion: General Theory and Speeding up BFGS Rules for Faster Second-Order Optimization

Robert Gower, Filip Hanzely, Peter Richtarik, Sebastian U. Stich

We present the first accelerated randomized algorithm for solving linear systems in Euclidean spaces. One essential problem of this type is the matrix inversion problem. In particular, our algorithm can be specialized to invert positive definite matrices in such a way that all iterates (approximate solutions) generated

by the algorithm are positive definite matrices themselves. This opens the way for many applications in the field of optimization and machine learning. As an application of our general theory, we develop the first accelerated (deterministic and stochastic) quasi-Newton updates. Our updates lead to provably more aggressive approximations of the inverse Hessian, and lead to speed-ups over classical non-accelerated rules in numerical experiments. Experiments with empirical risk minimization show that our rules can accelerate training of machine learning models.

Blind Deconvolutional Phase Retrieval via Convex Programming

Ali Ahmed, Alireza Aghasi, Paul Hand

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, Jure Leskovec

Generating novel graph structures that optimize given objectives while obeying some given underlying rules is fundamental for chemistry, biology and social science research. This is especially important in the task of molecular graph generation, whose goal is to discover novel molecules with desired properties such as drug-likeness and synthetic accessibility, while obeying physical laws such as chemical valency. However, designing models that finds molecules that optimize desired properties while incorporating highly complex and non-differentiable rules remains to be a challenging task. Here we propose Graph Convolutional Policy Network (GCPN), a general graph convolutional network based model for goal-directed graph generation through reinforcement learning. The model is trained to optimize domain-specific rewards and adversarial loss through policy gradient, and acts in an environment that incorporates domain-specific rules. Experimental results show that GCPN can achieve 61% improvement on chemical property optimization over state-of-the-art baselines while resembling known molecules, and achieve 18.4% improvement on the constrained property optimization task.

Spectral Filtering for General Linear Dynamical Systems

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.

Dialog-to-Action: Conversational Question Answering Over a Large-Scale Knowledge Base

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.

Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language

Seonghyeon Nam, Yunji Kim, Seon Joo Kim

This paper addresses the problem of manipulating images using natural language description. Our task aims to semantically modify visual attributes of an object in an image according to the text describing the new visual appearance. Although existing methods synthesize images having new attributes, they do not fully preserve text-irrelevant contents of the original image. In this paper, we propose the text-adaptive generative adversarial network (TAGAN) to generate semantically manipulated images while preserving text-irrelevant contents. The key to our method is the text-adaptive discriminator that creates word level local discriminators according to input text to classify fine-grained attributes independently. With this discriminator, the generator learns to generate images where only regions that correspond to the given text is modified. Experimental results show that our method outperforms existing methods on CUB and Oxford-102 datasets, and our results were mostly preferred on a user study. Extensive analysis shows that our method is able to effectively disentangle visual attributes and produce pleasing outputs.

Exploration in Structured Reinforcement Learning

Jungseul Ok, Alexandre Proutiere, Damianos Tranos

We address reinforcement learning problems with finite state and action spaces where the underlying MDP has some known structure that could be potentially exploited to minimize the exploration rates of suboptimal (state, action) pairs. For any arbitrary structure, we derive problem-specific regret lower bounds satisfied by any learning algorithm. These lower bounds are made explicit for unstructured MDPs and for those whose transition probabilities and average reward functions are Lipschitz continuous w.r.t. the state and action. For Lipschitz MDPs, the bounds are shown not to scale with the sizes S and A of the state and action spaces, i.e., they are smaller than $c \log T$ where T is the time horizon and the constant c only depends on the Lipschitz structure, the span of the bias function, and the minimal action sub-optimality gap. This contrasts with unstructured MDPs where the regret lower bound typically scales as $SA \log T$. We devise DEL (Directed Exploration Learning), an algorithm that matches our regret lower bounds. We further simplify the algorithm for Lipschitz MDPs, and show that the simplified version is still able to efficiently exploit the structure.

Tree-to-tree Neural Networks for Program Translation

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. Furthermore, 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.

Virtual Class Enhanced Discriminative Embedding Learning

Binghui Chen, Weihong Deng, Haifeng Shen

Recently, learning discriminative features to improve the recognition performances gradually becomes the primary goal of deep learning, and numerous remarkable works have emerged. In this paper, we propose a novel yet extremely simple method Virtual Softmax to enhance the discriminative property of learned features by injecting a dynamic virtual negative class into the original softmax. Injecting

virtual class aims to enlarge inter-class margin and compress intra-class distribution by strengthening the decision boundary constraint. Although it seems weird to optimize with this additional virtual class, we show that our method derives from an intuitive and clear motivation, and it indeed encourages the features to be more compact and separable. This paper empirically and experimentally demonstrates the superiority of Virtual Softmax, improving the performances on a variety of object classification and face verification tasks.

Neural Networks Trained to Solve Differential Equations Learn General Representations

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 complementary to, another study of intrinsic dimensionality.

Deep Generative Models with Learnable Knowledge Constraints

Zhiting Hu, Zichao Yang, Russ R. Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, Eric P. Xing

The broad set of deep generative models (DGMs) has achieved remarkable advances. However, it is often difficult to incorporate rich structured domain knowledge with the end-to-end DGMs. Posterior regularization (PR) offers a principled framework to impose structured constraints on probabilistic models, but has limited applicability to the diverse DGMs that can lack a Bayesian formulation or even explicit density evaluation. PR also requires constraints to be fully specified $\{\text{a priori}\}$, which is impractical or suboptimal for complex knowledge with learnable uncertain parts. In this paper, we establish mathematical correspondence between PR and reinforcement learning (RL), and, based on the connection, expand PR to learn constraints as the extrinsic reward in RL. The resulting algorithm is model-agnostic to apply to any DGMs, and is flexible to adapt arbitrary constraints with the model jointly. Experiments on human image generation and templated sentence generation show models with learned knowledge constraints by our algorithm greatly improve over base generative models.

How to Start Training: The Effect of Initialization and Architecture

Boris Hanin, David Rolnick

We identify and study two common failure modes for early training in deep ReLU nets. For each, we give a rigorous proof of when it occurs and how to avoid it, for fully connected, convolutional, and residual architectures. We show that the first failure mode, exploding or vanishing mean activation length, can be avoided by initializing weights from a symmetric distribution with variance $2/\text{fan-in}$ and, for ResNets, by correctly scaling the residual modules. We prove that the second failure mode, exponentially large variance of activation length, never occurs in residual nets once the first failure mode is avoided. In contrast, for fully connected nets, we prove that this failure mode can happen and is avoided by keeping constant the sum of the reciprocals of layer widths. We demonstrate empirically the effectiveness of our theoretical results in predicting when networks are able to start training. In particular, we note that many popular initializations fail our criteria, whereas correct initialization and architecture allows much deeper networks to be trained.

Video-to-Video Synthesis

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, 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.)

Near-Optimal Policies for Dynamic Multinomial Logit Assortment Selection Models

Yining Wang, Xi Chen, Yuan Zhou

In this paper we consider the dynamic assortment selection problem under an uncapacitated multinomial-logit (MNL) model. By carefully analyzing a revenue potential function, we show that a trisection based algorithm achieves an item-independent regret bound of $O(\sqrt{T \log \log T})$, which matches information theoretical lower bounds up to iterated logarithmic terms. Our proof technique draws tools from the unimodal/convex bandit literature as well as adaptive confidence parameters in minimax multi-armed bandit problems.

Occam's razor is insufficient to infer the preferences of irrational agents

Stuart Armstrong, Sören Mindermann

Inverse reinforcement learning (IRL) attempts to infer human rewards or preferences from observed behavior. Since human planning systematically deviates from rationality, several approaches have been tried to account for specific human shortcomings.

However, the general problem of inferring the reward function of an agent of unknown rationality has received little attention.

Unlike the well-known ambiguity problems in IRL, this one is practically relevant but cannot be resolved by observing the agent's policy in enough environments. This paper shows (1) that a No Free Lunch result implies it is impossible to uniquely decompose a policy into a planning algorithm and reward function, and (2) that even with a reasonable simplicity prior/Occam's razor on the set of decompositions, we cannot distinguish between the true decomposition and others that lead to high regret.

To address this, we need simple 'normative' assumptions, which cannot be deduced exclusively from observations.

Bandit Learning with Implicit Feedback

Yi Qi, Qingyun Wu, Hongning Wang, Jie Tang, Maosong Sun

Implicit feedback, such as user clicks, although abundant in online information service systems, does not provide substantial evidence on users' evaluation of system's output. Without proper modeling, such incomplete supervision inevitably misleads model estimation, especially in a bandit learning setting where the feedback is acquired on the fly. In this work, we perform contextual bandit learning with implicit feedback by modeling the feedback as a composition of user results

t examination and relevance judgment. Since users' examination behavior is unobserved, we introduce latent variables to model it. We perform Thompson sampling on top of variational Bayesian inference for arm selection and model update. Our upper regret bound analysis of the proposed algorithm proves its feasibility of learning from implicit feedback in a bandit setting; and extensive empirical evaluations on click logs collected from a major MOOC platform further demonstrate its learning effectiveness in practice.

Adversarial Regularizers in Inverse Problems

Sebastian Lunz, Ozan Öktem, Carola-Bibiane Schönlieb

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.

The emergence of multiple retinal cell types through efficient coding of natural movies

Samuel Ocko, Jack Lindsey, Surya Ganguli, Stephane Deny

One of the most striking aspects of early visual processing in the retina is the immediate parcellation of visual information into multiple parallel pathways, formed by different retinal ganglion cell types each tiling the entire visual field. Existing theories of efficient coding have been unable to account for the functional advantages of such cell-type diversity in encoding natural scenes. Here we go beyond previous theories to analyze how a simple linear retinal encoding model with different convolutional cell types efficiently encodes naturalistic spatiotemporal movies given a fixed firing rate budget. We find that optimizing the receptive fields and cell densities of two cell types makes them match the properties of the two main cell types in the primate retina, midget and parasol cells, in terms of spatial and temporal sensitivity, cell spacing, and their relative ratio. Moreover, our theory gives a precise account of how the ratio of midget to parasol cells decreases with retinal eccentricity. Also, we train a nonlinear encoding model with a rectifying nonlinearity to efficiently encode naturalistic movies, and again find emergent receptive fields resembling those of midget and parasol cells that are now further subdivided into ON and OFF types. Thus our work provides a theoretical justification, based on the efficient coding of natural movies, for the existence of the four most dominant cell types in the primate retina that together comprise 70% of all ganglion cells.

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

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 c

lass (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>

Dropping Symmetry for Fast Symmetric Nonnegative Matrix Factorization

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.

On the Local Minima of the Empirical Risk

Chi Jin, Lydia T. Liu, Rong Ge, Michael I. Jordan

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

GIANT: Globally Improved Approximate Newton Method for Distributed Optimization

Shusen Wang, Fred Roosta, Peng Xu, Michael W. Mahoney

For distributed computing environment, we consider the empirical risk minimization problem and propose a distributed and communication-efficient Newton-type optimization method. At every iteration, each worker locally finds an Approximate Newton (ANT) direction, which is sent to the main driver. The main driver, then, averages all the ANT directions received from workers to form a Globally Improved ANT (GIANT) direction. GIANT is highly communication efficient and naturally exploits the trade-offs between local computations and global communications in that more local computations result in fewer overall rounds of communications. Theoretically, we show that GIANT enjoys an improved convergence rate as compared with first-order methods and existing distributed Newton-type methods. Further, and in sharp contrast with many existing distributed Newton-type methods, as well as popular first-order methods, a highly advantageous practical feature of GIANT is that it only involves one tuning parameter. We conduct large-scale experiments on a computer cluster and, empirically, demonstrate the superior performance of GIANT.

Stochastic Cubic Regularization for Fast Nonconvex Optimization

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, Michael I. Jordan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Derivative Estimation in Random Design

Yu Liu, Kris De Brabanter

We propose a nonparametric derivative estimation method for random design without

having to estimate the regression function. The method is based on a variance-reducing linear combination of symmetric difference quotients. First, we discuss the special case of uniform random design and establish the estimator's asymptotic

properties. Secondly, we generalize these results for any distribution of the dependent variable and compare the proposed estimator with popular estimators for derivative estimation such as local polynomial regression and smoothing splines.

Approximating Real-Time Recurrent Learning with Random Kronecker Factors

Asier Mujika, Florian Meier, Angelika Steger

Despite all the impressive advances of recurrent neural networks, sequential data is still in need of better modelling. Truncated backpropagation through time (TBPTT), the learning algorithm most widely used in practice, suffers from the truncation bias, which drastically limits its ability to learn long-term dependencies. The Real Time Recurrent Learning algorithm (RTRL) addresses this issue, but its high computational requirements make it infeasible in practice. The Unbiased Online Recurrent Optimization algorithm (UORO) approximates RTRL with a smaller runtime and memory cost, but with the disadvantage of obtaining noisy gradients that also limit its practical applicability. In this paper we propose the Kronecker Factored RTRL (KF-RTRL) algorithm that uses a Kronecker product decomposition to approximate the gradients for a large class of RNNs. We show that KF-RTRL is an unbiased and memory efficient online learning algorithm. Our theoretical analysis shows that, under reasonable assumptions, the noise introduced by our algorithm is not only stable over time but also asymptotically much smaller than the one of the UORO algorithm. We also confirm these theoretical results experimentally. Further, we show empirically that the KF-RTRL algorithm captures long-term dependencies and almost matches the performance of TBPTT on real world tasks by training Recurrent Highway Networks on a synthetic string memorization task and on the Penn TreeBank task, respectively. These results indicate that RTRL based approaches might be a promising future alternative to TBPTT.

Hyperbolic Neural Networks

Octavian Ganea, Gary Becigneul, Thomas Hofmann

Hyperbolic spaces have recently gained momentum in the context of machine learning due to their high capacity and tree-likeness properties. However, the representational power of hyperbolic geometry is not yet on par with Euclidean geometry, firstly because of the absence of corresponding hyperbolic neural network layers. Here, we bridge this gap in a principled manner by combining the formalism of Möbius gyrovector spaces with the Riemannian geometry of the Poincaré model of hyperbolic spaces. As a result, we derive hyperbolic versions of important deep learning tools: multinomial logistic regression, feed-forward and recurrent neural networks. This allows to embed sequential data and perform classification in the hyperbolic space. Empirically, we show that, even if hyperbolic optimization tools are limited, hyperbolic sentence embeddings either outperform or are on par with their Euclidean variants on textual entailment and noisy-prefix recognition tasks.

Mean Field for the Stochastic Blockmodel: Optimization Landscape and Convergence Issues

Soumendu Sundar Mukherjee, Purnamrita Sarkar, Y. X. Rachel Wang, Bowei Yan

Variational approximation has been widely used in large-scale Bayesian inference recently, the simplest kind of which involves imposing a mean field assumption to approximate complicated latent structures. Despite the computational scalability of mean field, theoretical studies of its loss function surface and the convergence behavior of iterative updates for optimizing the loss are far from complete. In this paper, we focus on the problem of community detection for a simple two-class Stochastic Blockmodel (SBM). Using batch co-ordinate ascent (BCAVI) for updates, we give a complete characterization of all the critical points and show different convergence behaviors with respect to initializations. When the parameters are known, we show a significant proportion of random initializations will converge to ground truth. On the other hand, when the parameters themselves need to be estimated, a random initialization will converge to an uninformative local optimum.

Fast Greedy MAP Inference for Determinantal Point Process to Improve Recommendation Diversity

Laming Chen, Guoxin Zhang, Eric Zhou

The determinantal point process (DPP) is an elegant probabilistic model of repulsion with applications in various machine learning tasks including summarization and search. However, the maximum a posteriori (MAP) inference for DPP which plays an important role in many applications is NP-hard, and even the popular greedy algorithm can still be too computationally expensive to be used in large-scale real-time scenarios. To overcome the computational challenge, in this paper, we propose a novel algorithm to greatly accelerate the greedy MAP inference for DPP. In addition, our algorithm also adapts to scenarios where the repulsion is only required among nearby few items in the result sequence. We apply the proposed algorithm to generate relevant and diverse recommendations. Experimental results show that our proposed algorithm is significantly faster than state-of-the-art competitors, and provides a better relevance-diversity trade-off on several public datasets, which is also confirmed in an online A/B test.

Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Andrea Vedaldi

While the use of bottom-up local operators in convolutional neural networks (CNNs) matches well some of the statistics of natural images, it may also prevent such models from capturing contextual long-range feature interactions. In this work, we propose a simple, lightweight approach for better context exploitation in CNNs. We do so by introducing a pair of operators: gather, which efficiently aggregates feature responses from a large spatial extent, and excite, which redistributes the pooled information to local features. The operators are cheap, both in terms of number of added parameters and computational complexity, and can be integrated directly in existing architectures to improve their performance. Experiments on several datasets show that gather-excite can bring benefits comparable to increasing the depth of a CNN at a fraction of the cost. For example, we find ResNet-50 with gather-excite operators is able to outperform its 101-layer counterpart on ImageNet with no additional learnable parameters. We also propose a parametric gather-excite operator pair which yields further performance gains, relate it to the recently-introduced Squeeze-and-Excitation Networks, and analyse the effects of these changes to the CNN feature activation statistics.

Active Learning for Non-Parametric Regression Using Purely Random Trees

Jack Goetz, Ambuj Tewari, Paul Zimmerman

Active learning is the task of using labelled data to select additional points to label, with the goal of fitting the most accurate model with a fixed budget of labelled points. In binary classification active learning is known to produce faster rates than passive learning for a broad range of settings. However in regression restrictive structure and tailored methods were previously needed to obtain theoretically superior performance. In this paper we propose an intuitive tree based active learning algorithm for non-parametric regression with provable improvement over random sampling. When implemented with Mondrian Trees our algorithm

hm is tuning parameter free, consistent and minimax optimal for Lipschitz functions.

DeepProbLog: Neural Probabilistic Logic Programming

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, Luc De Raedt

We introduce DeepProbLog, a probabilistic logic programming language that incorporates deep learning by means of neural predicates. We show how existing inference and learning techniques can be adapted for the new language. Our experiments demonstrate that DeepProbLog supports (i) both symbolic and subsymbolic representations and inference, (ii) program induction, (iii) probabilistic (logic) programming, and (iv) (deep) learning from examples. To the best of our knowledge, this work is the first to propose a framework where general-purpose neural networks and expressive probabilistic-logical modeling and reasoning are integrated in a way that exploits the full expressiveness and strengths of both worlds and can be trained end-to-end based on examples.

Image-to-image translation for cross-domain disentanglement

Abel Gonzalez-Garcia, Joost van de Weijer, Yoshua Bengio

Deep image translation methods have recently shown excellent results, outputting high-quality images covering multiple modes of the data distribution. There has also been increased interest in disentangling the internal representations learned by deep methods to further improve their performance and achieve a finer control. In this paper, we bridge these two objectives and introduce the concept of cross-domain disentanglement. We aim to separate the internal representation into three parts. The shared part contains information for both domains. The exclusive parts, on the other hand, contain only factors of variation that are particular to each domain. We achieve this through bidirectional image translation based on Generative Adversarial Networks and cross-domain autoencoders, a novel network component. Our model offers multiple advantages. We can output diverse samples covering multiple modes of the distributions of both domains, perform domain-specific image transfer and interpolation, and cross-domain retrieval without the need of labeled data, only paired images. We compare our model to the state-of-the-art in multi-modal image translation and achieve better results for translation on challenging datasets as well as for cross-domain retrieval on realistic datasets.

Expanding Holographic Embeddings for Knowledge Completion

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).

Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation

Qiang Liu, Lihong Li, Ziyang Tang, Dengyong Zhou

We consider the off-policy estimation problem of estimating the expected reward of a target policy using samples collected by a different behavior policy. Importance sampling (IS) has been a key technique to derive (nearly) unbiased estimates

ors, but is known to suffer from an excessively high variance in long-horizon problems. In the extreme case of infinite-horizon problems, the variance of an IS-based estimator may even be unbounded. In this paper, we propose a new off-policy estimation method that applies IS directly on the stationary state-visit distributions to avoid the exploding variance issue faced by existing estimators. Our key contribution is a novel approach to estimating the density ratio of two stationary distributions, with trajectories sampled from only the behavior distribution. We develop a mini-max loss function for the estimation problem, and derive a closed-form solution for the case of RKHS. We support our method with both theoretical and empirical analyses.

TopRank: A practical algorithm for online stochastic ranking

Tor Lattimore, Branislav Kveton, Shuai Li, Csaba Szepesvari

Online learning to rank is a sequential decision-making problem where in each round the learning agent chooses a list of items and receives feedback in the form of clicks from the user. Many sample-efficient algorithms have been proposed for this problem that assume a specific click model connecting rankings and user behavior. We propose a generalized click model that encompasses many existing models, including the position-based and cascade models. Our generalization motivates a novel online learning algorithm based on topological sort, which we call TopRank. TopRank is (a) more natural than existing algorithms, (b) has stronger regret guarantees than existing algorithms with comparable generality, (c) has a more insightful proof that leaves the door open to many generalizations, (d) outperforms existing algorithms empirically.

rho-POMDPs have Lipschitz-Continuous epsilon-Optimal Value Functions

Mathieu Fehr, Olivier Buffet, Vincent Thomas, Jilles Dibangoye

Many state-of-the-art algorithms for solving Partially Observable Markov Decision Processes (POMDPs) rely on turning the problem into a "fully observable" problem—a belief MDP—and exploiting the piece-wise linearity and convexity (PWLC) of the optimal value function in this new state space (the belief simplex Δ). This approach has been extended to solving ρ -POMDPs—i.e., for information-oriented criteria—when the reward ρ is convex in Δ . General ρ -POMDPs can also be turned into "fully observable" problems, but with no means to exploit the PWLC property. In this paper, we focus on POMDPs and ρ -POMDPs with $\lambda \rho$ -Lipschitz reward function, and demonstrate that, for finite horizons, the optimal value function is Lipschitz-continuous. Then, value function approximators are proposed for both upper- and lower-bounding the optimal value function, which are shown to provide uniformly improvable bounds. This allows proposing two algorithms derived from HSVI which are empirically evaluated on various benchmark problems.

Minimax Estimation of Neural Net Distance

Kaiyi Ji, Yingbin Liang

An important class of distance metrics proposed for training generative adversarial networks (GANs) is the integral probability metric (IPM), in which the neural net distance captures the practical GAN training via two neural networks. This paper investigates the minimax estimation problem of the neural net distance based on samples drawn from the distributions. We develop the first known minimax lower bound on the estimation error of the neural net distance, and an upper bound tighter than an existing bound on the estimator error for the empirical neural net distance. Our lower and upper bounds match not only in the order of the sample size but also in terms of the norm of the parameter matrices of neural networks, which justifies the empirical neural net distance as a good approximation of the true neural net distance for training GANs in practice.

Using Large Ensembles of Control Variates for Variational Inference

Tomas Geffner, Justin Domke

Variational inference is increasingly being addressed with stochastic optimization. In this setting, the gradient's variance plays a crucial role in the optimization procedure, since high variance gradients lead to poor convergence. A popular

an approach used to reduce gradient's variance involves the use of control variates. Despite the good results obtained, control variates developed for variational inference are typically looked at in isolation. In this paper we clarify the large number of control variates that are available by giving a systematic view of how they are derived. We also present a Bayesian risk minimization framework in which the quality of a procedure for combining control variates is quantified by its effect on optimization convergence rates, which leads to a very simple combination rule. Results show that combining a large number of control variates this way significantly improves the convergence of inference over using the typical gradient estimators or a reduced number of control variates.

Deep Generative Markov State Models

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.

Zero-Shot Transfer with Deictic Object-Oriented Representation in Reinforcement Learning

Ofir Marom, Benjamin Rosman

Object-oriented representations in reinforcement learning have shown promise in transfer learning, with previous research introducing a propositional object-oriented framework that has provably efficient learning bounds with respect to sample complexity. However, this framework has limitations in terms of the classes of tasks it can efficiently learn. In this paper we introduce a novel deictic object-oriented framework that has provably efficient learning bounds and can solve a broader range of tasks. Additionally, we show that this framework is capable of zero-shot transfer of transition dynamics across tasks and demonstrate this empirically for the Taxi and Sokoban domains.

Practical Methods for Graph Two-Sample Testing

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.

Point process latent variable models of larval zebrafish behavior

Anuj Sharma, Robert Johnson, Florian Engert, Scott Linderman

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.

Learning to Navigate in Cities Without a Map

Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, koray kavukcuoglu, Andrew Zisserman, Raiha Hadsell

Navigating through unstructured environments is a basic capability of intelligent creatures, and thus is of fundamental interest in the study and development of artificial intelligence. Long-range navigation is a complex cognitive task that relies on developing an internal representation of space, grounded by recognisable landmarks and robust visual processing, that can simultaneously support continuous self-localisation ("I am here") and a representation of the goal ("I am going there"). Building upon recent research that applies deep reinforcement learning to maze navigation problems, we present an end-to-end deep reinforcement learning approach that can be applied on a city scale. Recognising that successful navigation relies on integration of general policies with locale-specific knowledge, we propose a dual pathway architecture that allows locale-specific features to be encapsulated, while still enabling transfer to multiple cities. A key contribution of this paper is an interactive navigation environment that uses Google Street View for its photographic content and worldwide coverage. Our baselines demonstrate that deep reinforcement learning agents can learn to navigate in multiple cities and to traverse to target destinations that may be kilometres away. A video summarizing our research and showing the trained agent in diverse city environments as well as on the transfer task is available at: <https://sites.google.com/view/learn-navigate-cities-nips18>

Fast greedy algorithms for dictionary selection with generalized sparsity constraints

Kaito Fujii, Tasuku Soma

In dictionary selection, several atoms are selected from finite candidates that successfully approximate given data points in the sparse representation. We propose a novel efficient greedy algorithm for dictionary selection. Not only does our algorithm work much faster than the known methods, but it can also handle more complex sparsity constraints, such as average sparsity. Using numerical experiments, we show that our algorithm outperforms the known methods for dictionary selection, achieving competitive performances with dictionary learning algorithms in a smaller running time.

Acceleration through Optimistic No-Regret Dynamics

Jun-Kun Wang, Jacob D. Abernethy

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation

Yuan Li, Xiaodan Liang, Zhiting Hu, Eric P. 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.

Dimensionally Tight Bounds for Second-Order Hamiltonian Monte Carlo

Oren Mangoubi, Nisheeth Vishnoi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Invertibility of Convolutional Generative Networks from Partial Measurements

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.

Towards Robust Detection of Adversarial Examples

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.

Bayesian Model-Agnostic Meta-Learning

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 propo

sed 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.

Learning Signed Determinantal Point Processes through the Principal Minor Assignment Problem

Victor-Emmanuel Brunel

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Direct Estimation of Differences in Causal Graphs

Yuhao Wang, Chandler Squires, Anastasiya Belyaeva, Caroline Uhler

We consider the problem of estimating the differences between two causal directed acyclic graph (DAG) models with a shared topological order given i.i.d. samples from each model. This is of interest for example in genomics, where changes in the structure or edge weights of the underlying causal graphs reflect alterations in the gene regulatory networks. We here provide the first provably consistent method for directly estimating the differences in a pair of causal DAGs without separately learning two possibly large and dense DAG models and computing their difference. Our two-step algorithm first uses invariance tests between regression coefficients of the two data sets to estimate the skeleton of the difference graph and then orients some of the edges using invariance tests between regression residual variances. We demonstrate the properties of our method through a simulation study and apply it to the analysis of gene expression data from ovarian cancer and during T-cell activation.

A²-Nets: Double Attention Networks

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.

Sparse Attentive Backtracking: Temporal Credit Assignment Through Reminding

Nan Rosemary Ke, Anirudh Goyal ALIAS PARTH GOYAL, Olexa Bilaniuk, Jonathan Binas, Michael C. 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 re

current 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.

Ridge Regression and Provable Deterministic Ridge Leverage Score Sampling
Shannon McCurdy

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Dynamic Network Model from Partial Observations

Elahe Ghaleb, Baharan Mirzasoleiman, Radu Grosu, Jure Leskovec

Can evolving networks be inferred and modeled without directly observing their nodes and edges? In many applications, the edges of a dynamic network might not be observed, but one can observe the dynamics of stochastic cascading processes (e.g., information diffusion, virus propagation) occurring over the unobserved network. While there have been efforts to infer networks based on such data, providing a generative probabilistic model that is able to identify the underlying time-varying network remains an open question. Here we consider the problem of inferring generative dynamic network models based on network cascade diffusion data. We propose a novel framework for providing a non-parametric dynamic network model---based on a mixture of coupled hierarchical Dirichlet processes---based on data capturing cascade node infection times. Our approach allows us to infer the evolving community structure in networks and to obtain an explicit predictive distribution over the edges of the underlying network---including those that were not involved in transmission of any cascade, or are likely to appear in the future. We show the effectiveness of our approach using extensive experiments on synthetic as well as real-world networks.

Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate

Mikhail Belkin, Daniel J. Hsu, Partha Mitra

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Actor-Critic Policy Optimization in Partially Observable Multiagent Environments
Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Perolat, Karl Tuyls, Remi Munos, Michael Bowling

Optimization of parameterized policies for reinforcement learning (RL) is an important and challenging problem in artificial intelligence. Among the most common

approaches are algorithms based on gradient ascent of a score function representing discounted return. In this paper, we examine the role of these policy gradient and actor-critic algorithms in partially-observable multiagent environments.

We show several candidate policy update rules and relate them to a foundation of regret minimization and multiagent learning techniques for the one-shot and tabular cases, leading to previously unknown convergence guarantees. We apply our method to model-free multiagent reinforcement learning in adversarial sequential decision problems (zero-sum imperfect information games), using RL-style function approximation. We evaluate on commonly used benchmark Poker domains, showing performance against fixed policies and empirical convergence to approximate Nash equilibria in self-play with rates similar to or better than a baseline model-free algorithm for zero-sum games, without any domain-specific state space reductions.

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

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.

Relational recurrent neural networks

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.

Improved Expressivity Through Dendritic Neural Networks

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 effluent 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 stre

length 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 outperforming naive deep neural network architectures when tested on 121 classification tasks from the UCI machine learning repository.

DAGs with NO TEARS: Continuous Optimization for Structure Learning

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, Eric P. Xing

Estimating the structure of directed acyclic graphs (DAGs, also known as Bayesian networks) is a challenging problem since the search space of DAGs is combinatorial and scales superexponentially with the number of nodes. Existing approaches rely on various local heuristics for enforcing the acyclicity constraint. In this paper, we introduce a fundamentally different strategy: we formulate the structure learning problem as a purely continuous optimization problem over real matrices that avoids this combinatorial constraint entirely.

This is achieved by a novel characterization of acyclicity that is not only smooth but also exact. The resulting problem can be efficiently solved by standard numerical algorithms, which also makes implementation effortless. The proposed method outperforms existing ones, without imposing any structural assumptions on the graph such as bounded treewidth or in-degree.

Kalman Normalization: Normalizing Internal Representations Across Network Layers
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.

Connectionist Temporal Classification with Maximum Entropy Regularization

Hu Liu, Sheng Jin, Changshui Zhang

Connectionist Temporal Classification (CTC) is an objective function for end-to-end sequence learning, which adopts dynamic programming algorithms to directly learn the mapping between sequences. CTC has shown promising results in many sequence learning applications including speech recognition and scene text recognition. However, CTC tends to produce highly peaky and overconfident distributions, which is a symptom of overfitting. To remedy this, we propose a regularization method based on maximum conditional entropy which penalizes peaky distributions and encourages exploration. We also introduce an entropy-based pruning method to dramatically reduce the number of CTC feasible paths by ruling out unreasonable alignments. Experiments on scene text recognition show that our proposed methods consistently improve over the CTC baseline without the need to adjust training settings. Code has been made publicly available at: <https://github.com/liuhu-bigeye/enctc.crnn>.

Are GANs Created Equal? A Large-Scale Study

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}.

Recurrent Transformer Networks for Semantic Correspondence

Seungryong Kim, Stephen Lin, SANG RYUL JEON, Dongbo Min, Kwanghoon Sohn

We present recurrent transformer networks (RTNs) for obtaining dense correspondences between semantically similar images. Our networks accomplish this through an iterative process of estimating spatial transformations between the input images and using these transformations to generate aligned convolutional activations. By directly estimating the transformations between an image pair, rather than employing spatial transformer networks to independently normalize each individual image, we show that greater accuracy can be achieved. This process is conducted in a recursive manner to refine both the transformation estimates and the feature representations. In addition, a technique is presented for weakly-supervised training of RTNs that is based on a proposed classification loss. With RTNs, state-of-the-art performance is attained on several benchmarks for semantic correspondence.

Generative Neural Machine Translation

Harshil Shah, David Barber

We introduce Generative Neural Machine Translation (GNMT), a latent variable architecture which is designed to model the semantics of the source and target sentences. We modify an encoder-decoder translation model by adding a latent variable as a language agnostic representation which is encouraged to learn the meaning of the sentence. GNMT achieves competitive BLEU scores on pure translation tasks, and is superior when there are missing words in the source sentence. We augment the model to facilitate multilingual translation and semi-supervised learning without adding parameters. This framework significantly reduces overfitting when there is limited paired data available, and is effective for translating between pairs of languages not seen during training.

FRAGE: Frequency-Agnostic Word Representation

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.

sks, including word similarity, language modeling, machine translation and text classification. Results show that we achieve higher performance than the baselines in all tasks.

Variational Memory Encoder-Decoder

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.

Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding

Hajin Shim, Sung Ju Hwang, Eunho Yang

We consider the problem of active feature acquisition where the goal is to sequentially select the subset of features in order to achieve the maximum prediction performance in the most cost-effective way at test time. In this work, we formulate this active feature acquisition as a jointly learning problem of training both the classifier (environment) and the RL agent that decides either to stop and predict' or collect a new feature' at test time, in a cost-sensitive manner. We also introduce a novel encoding scheme to represent acquired subsets of features by proposing an order-invariant set encoding at the feature level, which also significantly reduces the search space for our agent. We evaluate our model on a carefully designed synthetic dataset for the active feature acquisition as well as several medical datasets. Our framework shows meaningful feature acquisition process for diagnosis that complies with human knowledge, and outperforms all baselines in terms of prediction performance as well as feature acquisition cost.

Data-Efficient Hierarchical Reinforcement Learning

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 substa

ntially outperforms previous state-of-the-art techniques.

Verifiable Reinforcement Learning via Policy Extraction

Osbert Bastani, Yewen Pu, Armando Solar-Lezama

While deep reinforcement learning has successfully solved many challenging control tasks, its real-world applicability has been limited by the inability to ensure the safety of learned policies. We propose an approach to verifiable reinforcement learning by training decision tree policies, which can represent complex policies (since they are nonparametric), yet can be efficiently verified using existing techniques (since they are highly structured). The challenge is that decision tree policies are difficult to train. We propose VIPER, an algorithm that combines ideas from model compression and imitation learning to learn decision tree policies guided by a DNN policy (called the oracle) and its Q-function, and show that it substantially outperforms two baselines. We use VIPER to (i) learn a provably robust decision tree policy for a variant of Atari Pong with a symbolic state space, (ii) learn a decision tree policy for a toy game based on Pong that provably never loses, and (iii) learn a provably stable decision tree policy for cart-pole. In each case, the decision tree policy achieves performance equal to that of the original DNN policy.

Stochastic Spectral and Conjugate Descent Methods

Dmitry Kovalev, Peter Richtarik, Eduard Gorbunov, Elnur Gasanov

The state-of-the-art methods for solving optimization problems in big dimensions are variants of randomized coordinate descent (RCD). In this paper we introduce a fundamentally new type of acceleration strategy for RCD based on the augmentation of the set of coordinate directions by a few spectral or conjugate directions. As we increase the number of extra directions to be sampled from, the rate of the method improves, and interpolates between the linear rate of RCD and a linear rate independent of the condition number. We develop and analyze also inexact variants of these methods where the spectral and conjugate directions are allowed to be approximate only. We motivate the above development by proving several negative results which highlight the limitations of RCD with importance sampling.

A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization

Zhize Li, Jian Li

We analyze stochastic gradient algorithms for optimizing nonconvex, nonsmooth finite-sum problems. In particular, the objective function is given by the summation of a differentiable (possibly nonconvex) component, together with a possibly non-differentiable but convex component.

We propose a proximal stochastic gradient algorithm based on variance reduction, called ProxSVRG+.

Our main contribution lies in the analysis of ProxSVRG+.

It recovers several existing convergence results and improves/generalizes them (in terms of the number of stochastic gradient oracle calls and proximal oracle calls).

In particular, ProxSVRG+ generalizes the best results given by the SCSG algorithm, recently proposed by [Lei et al., NIPS'17] for the smooth nonconvex case.

ProxSVRG+ is also more straightforward than SCSG and yields simpler analysis.

Moreover, ProxSVRG+ outperforms the deterministic proximal gradient descent (ProxGD) for a wide range of minibatch sizes, which partially solves an open problem proposed in [Reddi et al., NIPS'16].

Also, ProxSVRG+ uses much less proximal oracle calls than ProxSVRG [Reddi et al., NIPS'16].

Moreover, for nonconvex functions satisfied Polyak-Łojasiewicz condition, we prove that ProxSVRG+ achieves a global linear convergence rate without restart unlike ProxSVRG.

Thus, it can \emph{automatically} switch to the faster linear convergence in some regions as long as the objective function satisfies the PL condition locally i

n these regions.

Finally, we conduct several experiments and the experimental results are consistent with the theoretical results.

Hierarchical Graph Representation Learning with Differentiable Pooling

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.

Robust Detection of Adversarial Attacks by Modeling the Intrinsic Properties of Deep Neural Networks

Zhihao Zheng, Pengyu Hong

It has been shown that deep neural network (DNN) based classifiers are vulnerable to human-imperceptible adversarial perturbations which can cause DNN classifiers to output wrong predictions with high confidence. We propose an unsupervised learning approach to detect adversarial inputs without any knowledge of attackers. Our approach tries to capture the intrinsic properties of a DNN classifier and uses them to detect adversarial inputs. The intrinsic properties used in this study are the output distributions of the hidden neurons in a DNN classifier presented with natural images. Our approach can be easily applied to any DNN classifiers or combined with other defense strategy to improve robustness. Experimental results show that our approach demonstrates state-of-the-art robustness in defending black-box and gray-box attacks.

Removing the Feature Correlation Effect of Multiplicative Noise

Zijun Zhang, Yining Zhang, Zongpeng Li

Multiplicative noise, including dropout, is widely used to regularize deep neural networks (DNNs), and is shown to be effective in a wide range of architectures and tasks. From an information perspective, we consider injecting multiplicative noise into a DNN as training the network to solve the task with noisy information pathways, which leads to the observation that multiplicative noise tends to increase the correlation between features, so as to increase the signal-to-noise ratio of information pathways. However, high feature correlation is undesirable, as it increases redundancy in representations. In this work, we propose non-correlating multiplicative noise (NCMN), which exploits batch normalization to remove the correlation effect in a simple yet effective way. We show that NCMN significantly improves the performance of standard multiplicative noise on image classification tasks, providing a better alternative to dropout for batch-normalized networks. Additionally, we present a unified view of NCMN and shake-shake regularization, which explains the performance gain of the latter.

The Effect of Network Width on the Performance of Large-batch Training

Lingjiao Chen, Hongyi Wang, Jinman Zhao, Dimitris Papailiopoulos, Paraschos Koutis

Distributed implementations of mini-batch stochastic gradient descent (SGD) suffer from communication overheads, attributed to the high frequency of gradient u

updates inherent in small-batch training. Training with large batches can reduce these overheads; however it besets the convergence of the algorithm and the generalization performance.

Efficient Loss-Based Decoding on Graphs for Extreme Classification

Itay Evron, Edward Moroshko, Koby Crammer

In extreme classification problems, learning algorithms are required to map instances to labels from an extremely large label set.

We build on a recent extreme classification framework with logarithmic time and space (LTLS), and on a general approach for error correcting output coding (ECOC) with loss-based decoding, and introduce a flexible and efficient approach accompanied by theoretical bounds.

Our framework employs output codes induced by graphs, for which we show how to perform efficient loss-based decoding to potentially improve accuracy.

In addition, our framework offers a tradeoff between accuracy, model size and prediction time.

We show how to find the sweet spot of this tradeoff using only the training data.

Our experimental study demonstrates the validity of our assumptions and claims, and shows that our method is competitive with state-of-the-art algorithms.

Scalable methods for 8-bit training of neural networks

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.

Solving Large Sequential Games with the Excessive Gap Technique

Christian Kroer, Gabriele Farina, Tuomas Sandholm

There has been tremendous recent progress on equilibrium-finding algorithms for zero-sum imperfect-information extensive-form games, but there has been a puzzling gap between theory and practice. First-order methods have significantly better theoretical convergence rates than any counterfactual-regret minimization (CFR) variant. Despite this, CFR variants have been favored in practice. Experiments with first-order methods have only been conducted on small- and medium-sized games because those methods are complicated to implement in this setting, and because CFR variants have been enhanced extensively for over a decade they perform well in practice. In this paper we show that a particular first-order method, a state-of-the-art variant of the excessive gap technique---instantiated with the dilated entropy distance function---can efficiently solve large real-world problems competitively with CFR and its variants. We show this on large endgames encountered by the Libratus poker AI, which recently beat top human poker specialist professionals at no-limit Texas hold'em. We show experimental results on our variant of the excessive gap technique as well as a prior version. We introduce a numerically friendly implementation of the smoothed best response computation ass

ociated with first-order methods for extensive-form game solving. We present, to our knowledge, the first GPU implementation of a first-order method for extensive-form games. We present comparisons of several excessive gap technique and CFR variants.

Step Size Matters in Deep Learning

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.

A Reduction for Efficient LDA Topic Reconstruction

Matteo Almanza, Flavio Chierichetti, Alessandro Panconesi, Andrea Vattani

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Convex Elicitation of Continuous Properties

Jessica Finocchiaro, Rafael Frongillo

A property or statistic of a distribution is said to be elicitable if it can be expressed as the minimizer of some loss function in expectation. Recent work shows that continuous real-valued properties are elicitable if and only if they are identifiable, meaning the set of distributions with the same property value can be described by linear constraints. From a practical standpoint, one may ask for which such properties do there exist convex loss functions. In this paper, in a finite-outcome setting, we show that in fact every elicitable real-valued property can be elicited by a convex loss function. Our proof is constructive, and leads to convex loss functions for new properties.

The Nearest Neighbor Information Estimator is Adaptively Near Minimax Rate-Optimal

Jiantao Jiao, Weihao Gao, Yanjun Han

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Bandit Learning with Positive Externalities

Virag Shah, Jose Blanchet, Ramesh Johari

In many platforms, user arrivals exhibit a self-reinforcing behavior: future user arrivals are likely to have preferences similar to users who were satisfied in the past. In other words, arrivals exhibit $\{\text{positive externalities}\}$. We study multiarmed bandit (MAB) problems with positive externalities. We show that the self-reinforcing preferences may lead standard benchmark algorithms such as UCB to exhibit linear regret. We develop a new algorithm, Balanced Exploration (BE), which explores arms carefully to avoid suboptimal convergence of arrivals before sufficient evidence is gathered. We also introduce an adaptive variant of BE

which successively eliminates suboptimal arms. We analyze their asymptotic regret, and establish optimality by showing that no algorithm can perform better.

Minimax Statistical Learning with Wasserstein distances

Jaeho Lee, Maxim Raginsky

As opposed to standard empirical risk minimization (ERM), distributionally robust optimization aims to minimize the worst-case risk over a larger ambiguity set containing the original empirical distribution of the training data. In this work, we describe a minimax framework for statistical learning with ambiguity sets given by balls in Wasserstein space. In particular, we prove generalization bounds that involve the covering number properties of the original ERM problem. As an illustrative example, we provide generalization guarantees for transport-based domain adaptation problems where the Wasserstein distance between the source and target domain distributions can be reliably estimated from unlabeled samples.

Dirichlet belief networks for topic structure learning

He Zhao, Lan Du, Wray Buntine, Mingyuan Zhou

Recently, considerable research effort has been devoted to developing deep architectures for topic models to learn topic structures. Although several deep models have been proposed to learn better topic proportions of documents, how to leverage the benefits of deep structures for learning word distributions of topics has not yet been rigorously studied. Here we propose a new multi-layer generative process on word distributions of topics, where each layer consists of a set of topics and each topic is drawn from a mixture of the topics of the layer above. As the topics in all layers can be directly interpreted by words, the proposed model is able to discover interpretable topic hierarchies. As a self-contained module, our model can be flexibly adapted to different kinds of topic models to improve their modelling accuracy and interpretability. Extensive experiments on text corpora demonstrate the advantages of the proposed model.

Efficient Stochastic Gradient Hard Thresholding

Pan Zhou, Xiaotong Yuan, Jiashi Feng

Stochastic gradient hard thresholding methods have recently been shown to work favorably in solving large-scale empirical risk minimization problems under sparsity or rank constraint. Despite the improved iteration complexity over full gradient methods, the gradient evaluation and hard thresholding complexity of the existing stochastic algorithms usually scales linearly with data size, which could still be expensive when data is huge and the hard thresholding step could be as expensive as singular value decomposition in rank-constrained problems. To address these deficiencies, we propose an efficient hybrid stochastic gradient hard thresholding (HSG-HT) method that can be provably shown to have sample-size-independent gradient evaluation and hard thresholding complexity bounds. Specifically, we prove that the stochastic gradient evaluation complexity of HSG-HT scales linearly with inverse of sub-optimality and its hard thresholding complexity scales logarithmically. By applying the heavy ball acceleration technique, we further propose an accelerated variant of HSG-HT which can be shown to have improved factor dependence on restricted condition number. Numerical results confirm our theoretical affirmation and demonstrate the computational efficiency of the proposed methods.

Learning long-range spatial dependencies with horizontal gated recurrent units

Drew Linsley, Junkyung Kim, Vijay Veerabadrán, 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.

Tight Bounds for Collaborative PAC Learning via Multiplicative Weights

Jiecao Chen, Qin Zhang, Yuan Zhou

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

(Probably) Concave Graph Matching

Haggai Maron, Yaron Lipman

In this paper we address the graph matching problem. Following the recent works of \cite{zaslavskiy2009path,Vestner2017} we analyze and generalize the idea of concave relaxations. We introduce the concepts of \emph{conditionally concave} and \emph{probably conditionally concave} energies on polytopes and show that they encapsulate many instances of the graph matching problem, including matching Euclidean graphs and graphs on surfaces. We further prove that local minima of probably conditionally concave energies on general matching polytopes (\eg, doubly stochastic) are with high probability extreme points of the matching polytope (\eg, permutations).

HOUDINI: Lifelong Learning as Program Synthesis

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.

Video Prediction via Selective Sampling

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.

Manifold-tiling Localized Receptive Fields are Optimal in Similarity-preserving Neural Networks

Anirvan Sengupta, Cengiz Pehlevan, Mariano Tepper, Alexander Genkin, Dmitri Chklovskii

Many neurons in the brain, such as place cells in the rodent hippocampus, have localized receptive fields, i.e., they respond to a small neighborhood of stimulus space. What is the functional significance of such representations and how can they arise? Here, we propose that localized receptive fields emerge in similarity-preserving networks of rectifying neurons that learn low-dimensional manifold representations populated by sensory inputs. Numerical simulations of such networks on standard datasets yield manifold-tiling localized receptive fields. More generally, we show analytically that, for data lying on symmetric manifolds, optimal solutions of objectives, from which similarity-preserving networks are derived, have localized receptive fields. Therefore, nonnegative similarity-preserving mapping (NSM) implemented by neural networks can model representations of continuous manifolds in the brain.

Snap ML: A Hierarchical Framework for Machine Learning

Celestine Dünnér, 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.

Embedding Logical Queries on Knowledge Graphs

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 "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.

Parsimonious Bayesian deep networks

Mingyuan Zhou

Combining Bayesian nonparametrics and a forward model selection strategy, we construct parsimonious Bayesian deep networks (PBDNs) that infer capacity-regularized network architectures from the data and require neither cross-validation nor fine-tuning when training the model. One of the two essential components of a PBDN is the development of a special infinite-wide single-hidden-layer neural network, whose number of active hidden units can be inferred from the data. The other one is the construction of a greedy layer-wise learning algorithm that uses a forward model selection criterion to determine when to stop adding another hidden layer. We develop both Gibbs sampling and stochastic gradient descent based maximum a posteriori inference for PBDNs, providing state-of-the-art classification accuracy and interpretable data subtypes near the decision boundaries, while maintaining low computational complexity for out-of-sample prediction.

Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion

Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, Honglak Lee

There is growing interest in combining model-free and model-based approaches in reinforcement learning with the goal of achieving the high performance of model-free algorithms with low sample complexity. This is difficult because an imperfect dynamics model can degrade the performance of the learning algorithm, and in sufficiently complex environments, the dynamics model will always be imperfect. As a result, a key challenge is to combine model-based approaches with model-free learning in such a way that errors in the model do not degrade performance. We propose stochastic ensemble value expansion (STEVE), a novel model-based technique that addresses this issue. By dynamically interpolating between model rollouts of various horizon lengths, STEVE ensures that the model is only utilized when doing so does not introduce significant errors. Our approach outperforms model-free baselines on challenging continuous control benchmarks with an order-of-magnitude increase in sample efficiency.

Learning Versatile Filters for Efficient Convolutional Neural Networks

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.

Neural Nearest Neighbors Networks

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.

Learning latent variable structured prediction models with Gaussian perturbations

Kevin Bello, Jean Honorio

The standard margin-based structured prediction commonly uses a maximum loss over all possible structured outputs. The large-margin formulation including latent variables not only results in a non-convex formulation but also increases the search space by a factor of the size of the latent space. Recent work has proposed the use of the maximum loss over random structured outputs sampled independently from some proposal distribution, with theoretical guarantees. We extend this work by including latent variables. We study a new family of loss functions under Gaussian perturbations and analyze the effect of the latent space on the generalization bounds. We show that the non-convexity of learning with latent variables originates naturally, as it relates to a tight upper bound of the Gibbs decoder distortion with respect to the latent space. Finally, we provide a formulation using random samples and relaxations that produces a tighter upper bound of the Gibbs decoder distortion up to a statistical accuracy, which enables a polynomial time evaluation of the objective function. We illustrate the method with synthetic experiments and a computer vision application.

Differentially Private Change-Point Detection

Rachel Cummings, Sara Krehbiel, Yajun Mei, Rui Tuo, Wanrong Zhang

The change-point detection problem seeks to identify distributional changes at an unknown change-point k^* in a stream of data. This problem appears in many important practical settings involving personal data, including biosurveillance, fault detection, finance, signal detection, and security systems. The field of differential privacy offers data analysis tools that provide powerful worst-case privacy guarantees. We study the statistical problem of change-point problem through the lens of differential privacy. We give private algorithms for both online and offline change-point detection, analyze these algorithms theoretically, and then provide empirical validation of these results.

Proximal Graphical Event Models

Debarun Bhattacharjya, Dharmashankar Subramanian, Tian Gao

Event datasets include events that occur irregularly over the timeline and are prevalent in numerous domains. We introduce proximal graphical event models (PGEM) as a representation of such datasets. PGEMs belong to a broader family of models that characterize relationships between various types of events, where the rate of occurrence of an event type depends only on whether or not its parents have occurred in the most recent history. The main advantage over the state of the art models is that they are entirely data driven and do not require additional inputs from the user, which can require knowledge of the domain such as choice of basis functions or hyperparameters in graphical event models. We theoretically justify our learning of optimal windows for parental history and the choices of parental sets, and the algorithm are sound and complete in terms of parent structure.

cture learning. We present additional efficient heuristics for learning PGEMs from data, demonstrating their effectiveness on synthetic and real datasets.

Low-rank Interaction with Sparse Additive Effects Model for Large Data Frames

Geneviève Robin, Hoi-To Wai, Julie Josse, Olga Klopp, Eric Moulines

Many applications of machine learning involve the analysis of large data frames -- matrices collecting heterogeneous measurements (binary, numerical, counts, etc.) across samples -- with missing values. Low-rank models, as studied by Udell et al. (2016), are popular in this framework for tasks such as visualization, clustering and missing value imputation. Yet, available methods with statistical guarantees and efficient optimization do not allow explicit modeling of main additive effects such as row and column, or covariate effects. In this paper, we introduce a low-rank interaction and sparse additive effects (LORIS) model which combines matrix regression on a dictionary and low-rank design, to estimate main effects and interactions simultaneously. We provide statistical guarantees in the form of upper bounds on the estimation error of both components. Then, we introduce a mixed coordinate gradient descent (MCGD) method which provably converges sub-linearly to an optimal solution and is computationally efficient for large scale data sets. We show on simulated and survey data that the method has a clear advantage over current practices.

Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels

Zhilu Zhang, Mert Sabuncu

Deep neural networks (DNNs) have achieved tremendous success in a variety of applications across many disciplines. Yet, their superior performance comes with the expensive cost of requiring correctly annotated large-scale datasets. Moreover, due to DNNs' rich capacity, errors in training labels can hamper performance. To combat this problem, mean absolute error (MAE) has recently been proposed as a noise-robust alternative to the commonly-used categorical cross entropy (CCE) loss. However, as we show in this paper, MAE can perform poorly with DNNs and large-scale datasets. Here, we present a theoretically grounded set of noise-robust loss functions that can be seen as a generalization of MAE and CCE. Proposed loss functions can be readily applied with any existing DNN architecture and algorithm, while yielding good performance in a wide range of noisy label scenarios. We report results from experiments conducted with CIFAR-10, CIFAR-100 and FASHION-MNIST datasets and synthetically generated noisy labels.

Inference Aided Reinforcement Learning for Incentive Mechanism Design in Crowdsourcing

Zehong Hu, Yitao Liang, Jie Zhang, Zhao Li, Yang Liu

Incentive mechanisms for crowdsourcing are designed to incentivize financially self-interested workers to generate and report high-quality labels. Existing mechanisms are often developed as one-shot static solutions, assuming a certain level of knowledge about worker models (expertise levels, costs for exerting efforts, etc.). In this paper, we propose a novel inference aided reinforcement mechanism that acquires data sequentially and requires no such prior assumptions. Specifically, we first design a Gibbs sampling augmented Bayesian inference algorithm to estimate workers' labeling strategies from the collected labels at each step. Then we propose a reinforcement incentive learning (RIL) method, building on top of the above estimates, to uncover how workers respond to different payments. RIL dynamically determines the payment without accessing any ground-truth labels. We theoretically prove that RIL is able to incentivize rational workers to provide high-quality labels both at each step and in the long run. Empirical results show that our mechanism performs consistently well under both rational and non-fully rational (adaptive learning) worker models. Besides, the payments offered by RIL are more robust and have lower variances compared to existing one-shot mechanisms.

Fast and Effective Robustness Certification

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, Martin Vechev
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Online Structured Laplace Approximations for Overcoming Catastrophic Forgetting
Hippolyt Ritter, Aleksandar Botev, David Barber

We introduce the Kronecker factored online Laplace approximation for overcoming catastrophic forgetting in neural networks. The method is grounded in a Bayesian online learning framework, where we recursively approximate the posterior after every task with a Gaussian, leading to a quadratic penalty on changes to the weights. The Laplace approximation requires calculating the Hessian around a mode, which is typically intractable for modern architectures. In order to make our method scalable, we leverage recent block-diagonal Kronecker factored approximations to the curvature. Our algorithm achieves over 90% test accuracy across a sequence of 50 instantiations of the permuted MNIST dataset, substantially outperforming related methods for overcoming catastrophic forgetting.

Optimization over Continuous and Multi-dimensional Decisions with Observational Data

Dimitris Bertsimas, Christopher McCord

We consider the optimization of an uncertain objective over continuous and multi-dimensional decision spaces in problems in which we are only provided with observational data. We propose a novel algorithmic framework that is tractable, asymptotically consistent, and superior to comparable methods on example problems. Our approach leverages predictive machine learning methods and incorporates information on the uncertainty of the predicted outcomes for the purpose of prescribing decisions. We demonstrate the efficacy of our method on examples involving both synthetic and real data sets.

Neural Architecture Search with Bayesian Optimisation and Optimal Transport
Kirthivasan Kandasamy, Willie Neiswanger, Jeff Schneider, Barnabas Poczos, Eric P. 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.

An Information-Theoretic Analysis for Thompson Sampling with Many Actions

Shi Dong, Benjamin Van Roy

Information-theoretic Bayesian regret bounds of Russo and Van Roy capture the dependence of regret on prior uncertainty. However, this dependence is through entropy, which can become arbitrarily large as the number of actions increases. We establish new bounds that depend instead on a notion of rate-distortion. Among other things, this allows us to recover through information-theoretic arguments

a near-optimal bound for the linear bandit. We also offer a bound for the logistic bandit that dramatically improves on the best previously available, though this bound depends on an information-theoretic statistic that we have only been able to quantify via computation.

BML: A High-performance, Low-cost Gradient Synchronization Algorithm for DML Training

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%.

Proximal SCOPE for Distributed Sparse Learning

Shenyi Zhao, Gong-Duo Zhang, Ming-Wei Li, Wu-Jun Li

Distributed sparse learning with a cluster of multiple machines has attracted much attention in machine learning, especially for large-scale applications with high-dimensional data. One popular way to implement sparse learning is to use L1 regularization. In this paper, we propose a novel method, called proximal SCOPE (pSCOPE), for distributed sparse learning with L1 regularization. pSCOPE is based on a cooperative autonomous local learning (CALL) framework. In the CALL framework of pSCOPE, we find that the data partition affects the convergence of the learning procedure, and subsequently we define a metric to measure the goodness of a data partition. Based on the defined metric, we theoretically prove that pSCOPE is convergent with a linear convergence rate if the data partition is good enough. We also prove that better data partition implies faster convergence rate. Furthermore, pSCOPE is also communication efficient. Experimental results on real data sets show that pSCOPE can outperform other state-of-the-art distributed methods for sparse learning.

BinGAN: Learning Compact Binary Descriptors with a Regularized GAN

Maciej Zieba, Piotr Sembercki, 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.

Incorporating Context into Language Encoding Models for fMRI

Shailee Jain, Alexander Huth

Language encoding models help explain language processing in the human brain by learning functions that predict brain responses from the language stimuli that elicited them. Current word embedding-based approaches treat each stimulus word independently and thus ignore the influence of context on language understanding.

In this work we instead build encoding models using rich contextual representat

ions derived from an LSTM language model. Our models show a significant improvement in encoding performance relative to state-of-the-art embeddings in nearly every brain area. By varying the amount of context used in the models and providing the models with distorted context, we show that this improvement is due to a combination of better word embeddings learned by the LSTM language model and contextual information. We are also able to use our models to map context sensitivity across the cortex. These results suggest that LSTM language models learn high-level representations that are related to representations in the human brain.

Communication Efficient Parallel Algorithms for Optimization on Manifolds

Bayan Sapatbayeva, Michael Zhang, Lizhen Lin

The last decade has witnessed an explosion in the development of models, theory and computational algorithms for ``big data'' analysis. In particular, distributed inference has served as a natural and dominating paradigm for statistical inference. However, the existing literature on parallel inference almost exclusively focuses on Euclidean data and parameters. While this assumption is valid for many applications, it is increasingly more common to encounter problems where the data or the parameters lie on a non-Euclidean space, like a manifold for example. Our work aims to fill a critical gap in the literature by generalizing parallel inference algorithms to optimization on manifolds. We show that our proposed algorithm is both communication efficient and carries theoretical convergence guarantees. In addition, we demonstrate the performance of our algorithm to the estimation of Fréchet means on simulated spherical data and the low-rank matrix completion problem over Grassmann manifolds applied to the Netflix prize data set.

Memory Augmented Policy Optimization for Program Synthesis and Semantic Parsing

Chen Liang, Mohammad Norouzi, Jonathan Berant, Quoc V. Le, Ni Lao

We present Memory Augmented Policy Optimization (MAPO), a simple and novel way to leverage a memory buffer of promising trajectories to reduce the variance of policy gradient estimate. MAPO is applicable to deterministic environments with discrete actions, such as structured prediction and combinatorial optimization tasks. We express the expected return objective as a weighted sum of two terms: an expectation over the high-reward trajectories inside the memory buffer, and a separate expectation over trajectories outside the buffer. To make an efficient algorithm of MAPO, we propose: (1) memory weight clipping to accelerate and stabilize training; (2) systematic exploration to discover high-reward trajectories; (3) distributed sampling from inside and outside of the memory buffer to scale up training. MAPO improves the sample efficiency and robustness of policy gradient, especially on tasks with sparse rewards. We evaluate MAPO on weakly supervised program synthesis from natural language (semantic parsing). On the WikiTableQuestions benchmark, we improve the state-of-the-art by 2.6%, achieving an accuracy of 46.3%. On the WikiSQL benchmark, MAPO achieves an accuracy of 74.9% with only weak supervision, outperforming several strong baselines with full supervision. Our source code is available at <https://goo.gl/TXBp4e>

Context-dependent upper-confidence bounds for directed exploration

Raksha Kumaraswamy, Matthew Schlegel, Adam White, Martha White

Directed exploration strategies for reinforcement learning are critical for learning an optimal policy in a minimal number of interactions with the environment.

Many algorithms use optimism to direct exploration, either through visitation estimates or upper confidence bounds, as opposed to data-inefficient strategies like ϵ -greedy that use random, undirected exploration. Most data-efficient exploration methods require significant computation, typically relying on a learned model to guide exploration. Least-squares methods have the potential to provide some of the data-efficiency benefits of model-based approaches—because they summarize past interactions—with the computation closer to that of model-free approaches. In this work, we provide a novel, computationally efficient, incremental exploration strategy, leveraging this property of least-squares temporal difference learning (LSTD). We derive upper confidence bounds on the action-values learned

by LSTD, with context-dependent (or state-dependent) noise variance. Such context-dependent noise focuses exploration on a subset of variable states, and allows for reduced exploration in other states. We empirically demonstrate that our algorithm can converge more quickly than other incremental exploration strategies using confidence estimates on action-values.

Adaptive Negative Curvature Descent with Applications in Non-convex Optimization
Mingrui Liu, Zhe Li, Xiaoyu Wang, Jinfeng Yi, Tianbao Yang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

LF-Net: Learning Local Features from Images

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.

Why so gloomy? A Bayesian explanation of human pessimism bias in the multi-armed bandit task

Dalin Guo, Angela J. Yu

How humans make repeated choices among options with imperfectly known reward outcomes is an important problem in psychology and neuroscience. This is often studied using multi-armed bandits, which is also frequently studied in machine learning. We present data from a human stationary bandit experiment, in which we vary the average abundance and variability of reward availability (mean and variance of reward rate distributions). Surprisingly, we find subjects significantly underestimate prior mean of reward rates -- based on their self-report, at the end of a game, on their reward expectation of non-chosen arms. Previously, human learning in the bandit task was found to be well captured by a Bayesian ideal learning model, the Dynamic Belief Model (DBM), albeit under an incorrect generative assumption of the temporal structure - humans assume reward rates can change over time even though they are actually fixed. We find that the "pessimism bias" in the bandit task is well captured by the prior mean of DBM when fitted to human choices; but it is poorly captured by the prior mean of the Fixed Belief Model (FBM), an alternative Bayesian model that (correctly) assumes reward rates to be constants. This pessimism bias is also incompletely captured by a simple reinforcement learning model (RL) commonly used in neuroscience and psychology, in terms of fitted initial Q-values. While it seems sub-optimal, and thus mysterious, that humans have an underestimated prior reward expectation, our simulations show that an underestimated prior mean helps to maximize long-term gain, if the observer assumes volatility when reward rates are stable and utilizes a softmax decision policy instead of the optimal one (obtainable by dynamic programming). This raises the intriguing possibility that the brain underestimates reward rates to compensate for the incorrect non-stationarity assumption in the generative model and a simplified decision policy.

CapProNet: Deep Feature Learning via Orthogonal Projections onto Capsule Subspaces

Liheng Zhang, Marzieh Edraki, Guo-Jun Qi

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Robust Subspace Approximation in a Stream

Roie Levin, Anish Prasad Sevekari, David Woodruff

We study robust subspace estimation in the streaming and distributed settings. Given a set of n data points $\{a_i\}_{i=1}^n$ in \mathbb{R}^d and an integer k , we wish to find a linear subspace S of dimension k for which $\sum_i M(\text{dist}(S, a_i))$ is minimized, where $\text{dist}(S, x) := \min\{y \in S\} \|x - y\|_2$, and $M()$ is some loss function. When M is the identity function, S gives a subspace that is more robust to outliers than that provided by the truncated SVD. Though the problem is NP-hard, it is approximable within a $(1+\epsilon)$ factor in polynomial time when k and ϵ are constant.

We give the first sublinear approximation algorithm for this problem in the turnstile streaming and arbitrary partition distributed models, achieving the same time guarantees as in the offline case. Our algorithm is the first based entirely on oblivious dimensionality reduction, and significantly simplifies prior methods for this problem, which held in neither the streaming nor distributed models.

PointCNN: Convolution On X-Transformed Points

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.

Learning convex bounds for linear quadratic control policy synthesis

Jack Umenberger, Thomas B. Schön

Learning to make decisions from observed data in dynamic environments remains a problem of fundamental importance in a number of fields, from artificial intelligence and robotics, to medicine and finance.

This paper concerns the problem of learning control policies for unknown linear dynamical systems so as to maximize a quadratic reward function.

We present a method to optimize the expected value of the reward over the posterior distribution of the unknown system parameters, given data.

The algorithm involves sequential convex programming, and enjoys reliable local convergence and robust stability guarantees.

Numerical simulations and stabilization of a real-world inverted pendulum are used to demonstrate the approach, with strong performance and robustness properties observed in both.

Manifold Structured Prediction

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. Specific

cally, 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.

Efficient Gradient Computation for Structured Output Learning with Rational and Tropical Losses

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, Dmitry Storcheus, Scott Yang

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Mean-field theory of graph neural networks in graph partitioning

Tatsuro Kawamoto, Masashi Tsubaki, Tomoyuki Obuchi

A theoretical performance analysis of the graph neural network (GNN) is presented. For classification tasks, the neural network approach has the advantage in terms of flexibility that it can be employed in a data-driven manner, whereas Bayesian inference requires the assumption of a specific model. A fundamental question is then whether GNN has a high accuracy in addition to this flexibility. Moreover, whether the achieved performance is predominately a result of the backpropagation or the architecture itself is a matter of considerable interest. To gain a better insight into these questions, a mean-field theory of a minimal GNN architecture is developed for the graph partitioning problem. This demonstrates a good agreement with numerical experiments.

Adversarially Robust Generalization Requires More Data

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 large sample complexity.

Assessing Generative Models via Precision and Recall

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, Sylvain Gelly

Recent advances in generative modeling have led to an increased interest in the study of statistical divergences as means of model comparison. Commonly used evaluation methods, such as the Frechet Inception Distance (FID), correlate well with the perceived quality of samples and are sensitive to mode dropping. However, these metrics are unable to distinguish between different failure cases since they only yield one-dimensional scores. We propose a novel definition of precision and recall for distributions which disentangles the divergence into two separate dimensions. The proposed notion is intuitive, retains desirable properties, and naturally leads to an efficient algorithm that can be used to evaluate generative models. We relate this notion to total variation as well as to recent evaluation metrics such as Inception Score and FID. To demonstrate the practical utility of the proposed approach we perform an empirical study on several variants of Generative Adversarial Networks and Variational Autoencoders. In an extensive set of experiments we show that the proposed metric is able to disentangle the q

uality of generated samples from the coverage of the target distribution.

Improved Network Robustness with Adversary Critic

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.

Fast deep reinforcement learning using online adjustments from the past

Steven Hansen, Alexander Pritzel, Pablo Sprechmann, Andre Barreto, Charles Blundell

We propose Ephemeral Value Adjustments (EVA): a means of allowing deep reinforcement learning agents to rapidly adapt to experience in their replay buffer.

EVA shifts the value predicted by a neural network with an estimate of the value function found by prioritised sweeping over experience tuples from the replay buffer near the current state. EVA combines a number of recent ideas around combining episodic memory-like structures into reinforcement learning agents: slot-based storage, content-based retrieval, and memory-based planning.

We show that EVA is performant on a demonstration task and Atari games.

A Practical Algorithm for Distributed Clustering and Outlier Detection

Jiecao Chen, Erfan Sadeqi Azer, Qin Zhang

We study the classic k-means/median clustering, which are fundamental problems in unsupervised learning, in the setting where data are partitioned across multiple sites, and where we are allowed to discard a small portion of the data by labeling them as outliers. We propose a simple approach based on constructing small summary for the original dataset. The proposed method is time and communication efficient, has good approximation guarantees, and can identify the global outliers effectively.

To the best of our knowledge, this is the first practical algorithm with theoretical guarantees for distributed clustering with outliers. Our experiments on both real and synthetic data have demonstrated the clear superiority of our algorithm against all the baseline algorithms in almost all metrics.

How Much Restricted Isometry is Needed In Nonconvex Matrix Recovery?

Richard Zhang, Cedric Jozs, Somayeh Sojoudi, Javad Lavaei

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Limited Memory Kelley's Method Converges for Composite Convex and Submodular Objectives

Song Zhou, Swati Gupta, Madeleine Udell

The original simplicial method (OSM), a variant of the classic Kelley's cutting plane method, has been shown to converge to the minimizer of a composite convex

and submodular objective, though no rate of convergence for this method was known. Moreover, OSM is required to solve subproblems in each iteration whose size grows linearly in the number of iterations. We propose a limited memory version of Kelley's method (L-KM) and of OSM that requires limited memory (at most $n+1$ constraints for an n -dimensional problem) independent of the iteration. We prove convergence for L-KM when the convex part of the objective g is strongly convex and show it converges linearly when g is also smooth. Our analysis relies on duality between minimization of the composite convex and submodular objective and minimization of a convex function over the submodular base polytope. We introduce a limited memory version, L-FCFW, of the Fully-Corrective Frank-Wolfe (FCFW) method with approximate correction, to solve the dual problem. We show that L-FCFW and L-KM are dual algorithms that produce the same sequence of iterates; hence both converge linearly (when g is smooth and strongly convex) and with limited memory. We propose L-KM to minimize composite convex and submodular objectives; however, our results on L-FCFW hold for general polytopes and may be of independent interest.

Towards Understanding Acceleration Tradeoff between Momentum and Asynchrony in Nonconvex Stochastic Optimization

Tianyi Liu, Shiyang Li, Jianping Shi, Enlu Zhou, Tuo Zhao

Asynchronous momentum stochastic gradient descent algorithms (Async-MSGD) have been widely used in distributed machine learning, e.g., training large collaborative filtering systems and deep neural networks. Due to current technical limit, however, establishing convergence properties of Async-MSGD for these highly complicated nonconvex problems is generally infeasible. Therefore, we propose to analyze the algorithm through a simpler but nontrivial nonconvex problem --- streaming PCA. This allows us to make progress toward understanding Async-MSGD and gaining new insights for more general problems. Specifically, by exploiting the diffusion approximation of stochastic optimization, we establish the asymptotic rate of convergence of Async-MSGD for streaming PCA. Our results indicate a fundamental tradeoff between asynchrony and momentum: To ensure convergence and acceleration through asynchrony, we have to reduce the momentum (compared with Sync-MSGD). To the best of our knowledge, this is the first theoretical attempt on understanding Async-MSGD for distributed nonconvex stochastic optimization. Numerical experiments on both streaming PCA and training deep neural networks are provided to support our findings for Async-MSGD.

Metric on Nonlinear Dynamical Systems with Perron-Frobenius Operators

Isao Ishikawa, Keisuke Fujii, Masahiro Ikeda, Yuka Hashimoto, Yoshinobu Kawahara

The development of a metric for structural data is a long-term problem in pattern recognition and machine learning. In this paper, we develop a general metric for comparing nonlinear dynamical systems that is defined with Perron-Frobenius operators in reproducing kernel Hilbert spaces. Our metric includes the existing fundamental metrics for dynamical systems, which are basically defined with principal angles between some appropriately-chosen subspaces, as its special cases. We also describe the estimation of our metric from finite data. We empirically illustrate our metric with an example of rotation dynamics in a unit disk in a complex plane, and evaluate the performance with real-world time-series data.

A Mathematical Model For Optimal Decisions In A Representative Democracy

Malik Magdon-Ismail, Lirong Xia

Direct democracy, where each voter casts one vote, fails when the average voter competence falls below 50%. This happens in noisy settings when voters have limited information. Representative democracy, where voters choose representatives to vote, can be an elixir in both these situations. We introduce a mathematical model for studying representative democracy, in particular understanding the parameters of a representative democracy that gives maximum decision making capability. Our main result states that under general and natural conditions,

Loss Functions for Multiset Prediction

Sean Welleck, Zixin Yao, Yu Gai, Jialin Mao, Zheng Zhang, Kyunghyun Cho

We study the problem of multiset prediction. The goal of multiset prediction is to train a predictor that maps an input to a multiset consisting of multiple items. Unlike existing problems in supervised learning, such as classification, ranking and sequence generation, there is no known order among items in a target multiset, and each item in the multiset may appear more than once, making this problem extremely challenging. In this paper, we propose a novel multiset loss function by viewing this problem from the perspective of sequential decision making.

The proposed multiset loss function is empirically evaluated on two families of datasets, one synthetic and the other real, with varying levels of difficulty, against various baseline loss functions including reinforcement learning, sequence, and aggregated distribution matching loss functions. The experiments reveal the effectiveness of the proposed loss function over the others.

SEGA: Variance Reduction via Gradient Sketching

Filip Hanzely, Konstantin Mishchenko, Peter Richtarik

We propose a novel randomized first order optimization method---SEGA (SkEtched GRAdient method)---which progressively throughout its iterations builds a variance-reduced estimate of the gradient from random linear measurements (sketches) of the gradient provided at each iteration by an oracle. In each iteration, SEGA updates the current estimate of the gradient through a sketch-and-project operation using the information provided by the latest sketch, and this is subsequently used to compute an unbiased estimate of the true gradient through a random relaxation procedure. This unbiased estimate is then used to perform a gradient step. Unlike standard subspace descent methods, such as coordinate descent, SEGA can be used for optimization problems with a non-separable proximal term. We provide a general convergence analysis and prove linear convergence for strongly convex objectives. In the special case of coordinate sketches, SEGA can be enhanced with various techniques such as importance sampling, minibatching and acceleration, and its rate is up to a small constant factor identical to the best-known rate of coordinate descent.

Sharp Bounds for Generalized Uniformity Testing

Ilias Diakonikolas, Daniel M. Kane, Alistair Stewart

We study the problem of generalized uniformity testing of a discrete probability distribution: Given samples from a probability distribution p over an unknown size discrete domain Ω , we want to distinguish, with probability at least $2/3$, between the case that p is uniform on some subset of Ω versus ϵ -far, in total variation distance, from any such uniform distribution. We establish tight bounds on the sample complexity of generalized uniformity testing. In more detail, we present a computationally efficient tester whose sample complexity is optimal, with in constant factors, and a matching worst-case information-theoretic lower bound. Specifically, we show that the sample complexity of generalized uniformity testing is $\Theta(1/(\epsilon^{4/3} ||p||_3) + 1/(\epsilon^2 ||p||_2))$.

Non-Local Recurrent Network for Image Restoration

Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, Thomas S. 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 neighbor

orhood 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.

Practical exact algorithm for trembling-hand equilibrium refinements in games
Gabriele Farina, Nicola Gatti, Tuomas Sandholm

Nash equilibrium strategies have the known weakness that they do not prescribe rational play in situations that are reached with zero probability according to the strategies themselves, for example, if players have made mistakes. Trembling-hand refinements---such as extensive-form perfect equilibria and quasi-perfect equilibria---remedy this problem in sound ways. Despite their appeal, they have not received attention in practice since no known algorithm for computing them scales beyond toy instances. In this paper, we design an exact polynomial-time algorithm for finding trembling-hand equilibria in zero-sum extensive-form games. It is several orders of magnitude faster than the best prior ones, numerically stable, and quickly solves game instances with tens of thousands of nodes in the game tree. It enables, for the first time, the use of trembling-hand refinements in practice.

Thermostat-assisted continuously-tempered Hamiltonian Monte Carlo for Bayesian learning

Rui Luo, Jianhong Wang, Yaodong Yang, Jun WANG, Zhanxing Zhu

In this paper, we propose a novel sampling method, the thermostat-assisted continuously-tempered Hamiltonian Monte Carlo, for the purpose of multimodal Bayesian learning. It simulates a noisy dynamical system by incorporating both a continuously-varying tempering variable and the Nos'e-Hoover thermostats. A significant benefit is that it is not only able to efficiently generate i.i.d. samples when the underlying posterior distributions are multimodal, but also capable of adaptively neutralising the noise arising from the use of mini-batches. While the properties of the approach have been studied using synthetic datasets, our experiments on three real datasets have also shown its performance gains over several strong baselines for Bayesian learning with various types of neural networks plugged in.

Flexible neural representation for physics prediction

Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li F. Fei-Fei, Josh Tenenbaum, Daniel L. 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.

A Stein variational Newton method

Gianluca Detommaso, Tiangang Cui, Youssef Marzouk, Alessio Spantini, Robert Schichl

Stein variational gradient descent (SVGD) was recently proposed as a general pur

pose nonparametric variational inference algorithm: it minimizes the Kullback-Leibler divergence between the target distribution and its approximation by implementing a form of functional gradient descent on a reproducing kernel Hilbert space [Liu & Wang, NIPS 2016]. In this paper, we accelerate and generalize the SVGD algorithm by including second-order information, thereby approximating a Newton-like iteration in function space. We also show how second-order information can lead to more effective choices of kernel. We observe significant computational gains over the original SVGD algorithm in multiple test cases.

Dual Swap Disentangling

Zunlei Feng, Xinchao Wang, Chenglong Ke, An-Xiang Zeng, Dacheng Tao, Mingli Song
Learning interpretable disentangled representations is a crucial yet challenging task. In this paper, we propose a weakly semi-supervised method, termed as Dual Swap Disentangling (DSD), for disentangling using both labeled and unlabeled data. Unlike conventional weakly supervised methods that rely on full annotations on the group of samples, we require only limited annotations on paired samples that indicate their shared attribute like the color. Our model takes the form of a dual autoencoder structure. To achieve disentangling using the labeled pairs, we follow a 'encoding-swap-decoding' process, where we first swap the parts of their encodings corresponding to the shared attribute, and then decode the obtained hybrid codes to reconstruct the original input pairs. For unlabeled pairs, we follow the 'encoding-swap-decoding' process twice on designated encoding parts and enforce the final outputs to approximate the input pairs. By isolating parts of the encoding and swapping them back and forth, we impose the dimension-wise modularity and portability of the encodings of the unlabeled samples, which implicitly encourages disentangling under the guidance of labeled pairs. This dual swap mechanism, tailored for semi-supervised setting, turns out to be very effective. Experiments on image datasets from a wide domain show that our model yields state-of-the-art disentangling performances.

Algorithmic Regularization in Learning Deep Homogeneous Models: Layers are Automatically Balanced

Simon S. Du, Wei Hu, Jason D. Lee

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Third-order Smoothness Helps: Faster Stochastic Optimization Algorithms for Finding Local Minima

Yaodong Yu, Pan Xu, Quanquan Gu

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues.

Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

Robot Learning in Homes: Improving Generalization and Reducing Dataset Bias

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.

LAG: Lazily Aggregated Gradient for Communication-Efficient Distributed Learning
Tianyi Chen, Georgios Giannakis, Tao Sun, Wotao Yin

This paper presents a new class of gradient methods for distributed machine learning that adaptively skip the gradient calculations to learn with reduced communication and computation. Simple rules are designed to detect slowly-varying gradients and, therefore, trigger the reuse of outdated gradients. The resultant gradient-based algorithms are termed Lazily Aggregated Gradient --- justifying our acronym LAG used henceforth. Theoretically, the merits of this contribution are: i) the convergence rate is the same as batch gradient descent in strongly-convex, convex, and nonconvex cases; and, ii) if the distributed datasets are heterogeneous (quantified by certain measurable constants), the communication rounds needed to achieve a targeted accuracy are reduced thanks to the adaptive reuse of lagged gradients. Numerical experiments on both synthetic and real data corroborate a significant communication reduction compared to alternatives.

Equality of Opportunity in Classification: A Causal Approach

Junzhe Zhang, Elias Bareinboim

The Equalized Odds (for short, EO) is one of the most popular measures of discrimination used in the supervised learning setting. It ascertains fairness through the balance of the misclassification rates (false positive and negative) across the protected groups -- e.g., in the context of law enforcement, an African-American defendant who would not commit a future crime will have an equal opportunity of being released, compared to a non-recidivating Caucasian defendant. Despite this noble goal, it has been acknowledged in the literature that statistical tests based on the EO are oblivious to the underlying causal mechanisms that generated the disparity in the first place (Hardt et al. 2016). This leads to a critical disconnect between statistical measures readable from the data and the meaning of discrimination in the legal system, where compelling evidence that the observed disparity is tied to a specific causal process deemed unfair by society is required to characterize discrimination. The goal of this paper is to develop a principled approach to connect the statistical disparities characterized by the EO and the underlying, elusive, and frequently unobserved, causal mechanisms that generated such inequality. We start by introducing a new family of counterfactual measures that allows one to explain the misclassification disparities in terms of the underlying mechanisms in an arbitrary, non-parametric structural causal model. This will, in turn, allow legal and data analysts to interpret currently deployed classifiers through causal lens, linking the statistical disparities found in the data to the corresponding causal processes. Leveraging the new family of counterfactual measures, we develop a learning procedure to construct a classifier that is statistically efficient, interpretable, and compatible with the basic human intuition of fairness. We demonstrate our results through experiments in both real (COMPAS) and synthetic datasets.

Modern Neural Networks Generalize on Small Data Sets

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 surp

risingly 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.
