Compositional Plan Vectors
Coline Devin, Daniel Geng, Pieter Abbeel, Trevor Darrell, Sergey Levine
Autonomous agents situated in real-world environments must be able to master large repertoires of skills.
While a single short skill can be learned quickly, it would be impractical to learn every task independently. Instead, the agent should share knowledge across behaviors such that each task can be learned efficiently, and such that the resulting model can generalize to new tasks, especially ones that are compositions or subsets of tasks seen previously.
A policy conditioned on a goal or demonstration has the potential to share knowledge between tasks if it sees enough diversity of inputs. However, these methods may not generalize to a more complex task at test time. We introduce compositional plan vectors (CPVs) to enable a policy to perform compositions of tasks without additional supervision. CPVs represent trajectories as the sum of the subtasks within them. We show that CPVs can be learned within a one-shot imitation learning framework without any additional supervision or information about task hierarchy, and enable a demonstration-conditioned policy to generalize to tasks that sequence twice as many skills as the tasks seen during training. Analogously to embeddings such as word2vec in NLP, CPVs can also support simple arithmetic operations -- for example, we can add the CPVs for two different tasks to command an agent to compose both tasks, without any additional training.
****
Learning to Propagate for Graph Meta-Learning
LU LIU, Tianyi Zhou, Guodong Long, Jing Jiang, Chengqi Zhang
Meta-learning extracts the common knowledge from learning different tasks and uses it for unseen tasks. It can signi■cantly improve tasks that suffer from insuf■cient training data, e.g., few-shot learning. In most meta-learning methods, tasks are implicitly related by sharing parameters or optimizer. In this paper, we show that a meta-learner that explicitly relates tasks on a graph describing the relations of their output dimensions (e.g., classes) can signi■cantly improve few-shot learning. The graph’s structure is usually free or cheap to obtain but has rarely been explored in previous works. We develop a novel meta-learner of this type for prototype based classi■cation, in which a prototype is generated for each class, such that the nearest neighbor search among the prototypes produces an accurate classi■cation. The meta-learner, called “Gated Propagation Network (GPN)”, learns to propagate messages between prototypes of different classes on the graph, so that learning the prototype of each class bene■ts from the data of other related classes. In GPN, an attention mechanism aggregates messages from neighboring classes of each class, with a gate choosing between the aggregated message and the message from the class itself. We train GPN on a sequence of tasks from many-shot to few-shot generated by subgraph sampling. During training, it is able to reuse and update previously achieved prototypes from the memory in a life-long learning cycle. In experiments, under different training-test discrepancy and test task generation settings, GPN outperforms recent meta-learning methods on two benchmark datasets. Code of GPN is publicly available at: https://github.com/liulu112601/Gated-Propagation-Net.
****
XNAS: Neural Architecture Search with Expert Advice
Niv Nayman, Asaf Noy, Tal Ridnik, Itamar Friedman, Rong Jin, Lihi Zelnik
This paper introduces a novel optimization method for differential neural architecture search, based on the theory of prediction with expert advice. Its optimization criterion is well fitted for an architecture-selection, i.e., it minimizes the regret incurred by a sub-optimal selection of operations.
Unlike previous search relaxations, that require hard pruning of architectures, our method is designed to dynamically wipe out inferior architectures and enhance superior ones.
It achieves an optimal worst-case regret bound and suggests the use of multiple learning-rates, based on the amount of information carried by the backward gradients.
Experiments show that our algorithm achieves a strong performance over several i

mage classification datasets.
Specifically, it obtains an error rate of 1.6% for CIFAR-10, 23.9% for ImageNet under mobile settings, and achieves state-of-the-art results on three additional datasets.
****
Multi-resolution Multi-task Gaussian Processes
Oliver Hamelijnck, Theodoros Damoulas, Kangrui Wang, Mark Girolami
We consider evidence integration from potentially dependent observation processes under varying spatio-temporal sampling resolutions and noise levels. We offer a multi-resolution multi-task (MRGP) framework that allows for both inter-task and intra-task multi-resolution and multi-fidelity. We develop shallow Gaussian Process (GP) mixtures that approximate the difficult to estimate joint likelihood with a composite one and deep GP constructions that naturally handle biases. In doing so, we generalize existing approaches and offer information-theoretic corrections and efficient variational approximations. We demonstrate the competitiveness of MRGPs on synthetic settings and on the challenging problem of hyper-local estimation of air pollution levels across London from multiple sensing modalities operating at disparate spatio-temporal resolutions.
****
Deep Equilibrium Models
Shaojie Bai, J. Zico Kolter, Vladlen Koltun
We present a new approach to modeling sequential data: the deep equilibrium model (DEQ). Motivated by an observation that the hidden layers of many existing deep sequence models converge towards some fixed point, we propose the DEQ approach that directly finds these equilibrium points via root-finding. Such a method is equivalent to running an infinite depth (weight-tied) feedforward network, but has the notable advantage that we can analytically backpropagate through the equilibrium point using implicit differentiation. Using this approach, training and prediction in these networks require only constant memory, regardless of the effective “depth” of the network. We demonstrate how DEQs can be applied to two state-of-the-art deep sequence models: self-attention transformers and trellis networks. On large-scale language modeling tasks, such as the WikiText-103 benchmark, we show that DEQs 1) often improve performance over these state-of-the-art models (for similar parameter counts); 2) have similar computational requirements to existing models; and 3) vastly reduce memory consumption (often the bottleneck for training large sequence models), demonstrating an up-to 88% memory reduction in our experiments. The code is available at https://github.com/locuslab/deq.
****
Cross Attention Network for Few-shot Classification
Ruibing Hou, Hong Chang, Bingpeng MA, Shiguang Shan, Xilin Chen
Few-shot classification aims to recognize unlabeled samples from unseen classes given only few labeled samples. The unseen classes and low-data problem make few-shot classification very challenging. Many existing approaches extracted features from labeled and unlabeled samples independently, as a result, the features are not discriminative enough. In this work, we propose a novel Cross Attention Network to address the challenging problems in few-shot classification. Firstly, Cross Attention Module is introduced to deal with the problem of unseen classes. The module generates cross attention maps for each pair of class feature and query sample feature so as to highlight the target object regions, making the extracted feature more discriminative. Secondly, a transductive inference algorithm is proposed to alleviate the low-data problem, which iteratively utilizes the unlabeled query set to augment the support set, thereby making the class features more representative. Extensive experiments on two benchmarks show our method is a simple, effective and computationally efficient framework and outperforms the state-of-the-arts.
****
Order Optimal One-Shot Distributed Learning
Arsalan Sharifnassab, Saber Salehkaleybar, S. Jamaloddin Golestani
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 Gaussian Processes on a Million Data Points
Ke Wang, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q. Weinberger, Andrew Gordon 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 prior to requesting a name change in the electronic proceedings.
****
Asymmetric Valleys: Beyond Sharp and Flat Local Minima
Haowei He, Gao Huang, Yang Yuan
Despite the non-convex nature of their loss functions, deep neural networks are known to generalize well when optimized with stochastic gradient descent (SGD). Recent work conjectures that SGD with proper con■guration is able to ■nd wide and ■at local minima, which are correlated with good generalization performance. In this paper, we observe that local minima of modern deep networks are more than being ■at or sharp. Instead, at a local minimum there exist many asymmetric directions such that the loss increases abruptly along one side, and slowly along the opposite side – we formally de■ne such minima as asymmetric valleys. Under mild assumptions, we ■rst prove that for asymmetric valleys, a solution biased towards the ■at side generalizes better than the exact empirical minimizer. Then, we show that performing weight averaging along the SGD trajectory implicitly induces such biased solutions. This provides theoretical explanations for a series of intriguing phenomena observed in recent work [25, 5, 51]. Finally, extensive empirical experiments on both modern deep networks and simple 2 layer networks are conducted to validate our assumptions and analyze the intriguing properties of asymmetric valleys.
****
Calculating Optimistic Likelihoods Using (Geodesically) Convex Optimization
Viet Anh Nguyen, Soroosh Shafieezadeh Abadeh, Man-Chung Yue, Daniel Kuhn, Wolfram Wiesemann
A fundamental problem arising in many areas of machine learning is the evaluation of the likelihood of a given observation under different nominal distributions. Frequently, these nominal distributions are themselves estimated from data, which makes them susceptible to estimation errors. We thus propose to replace each nominal distribution with an ambiguity set containing all distributions in its vicinity and to evaluate an optimistic likelihood, that is, the maximum of the likelihood over all distributions in the ambiguity set. When the proximity of distributions is quantified by the Fisher-Rao distance or the Kullback-Leibler divergence, the emerging optimistic likelihoods can be computed efficiently using either geodesic or standard convex optimization techniques. We showcase the advantages of working with optimistic likelihoods on a classification problem using synthetic as well as empirical data.
****
Think out of the "Box": Generically-Constrained Asynchronous Composite Optimization and Hedging
Pooria Joulani, András György, Csaba Szepesvari
We present two new algorithms, ASYNCADA and HEDGEHOG, for asynchronous sparse online and stochastic optimization. ASYNCADA is, to our knowledge, the first asynchronous stochastic optimization algorithm with finite-time data-dependent convergence guarantees for generic convex constraints. In addition, ASYNCADA: (a) allows for proximal (i.e., composite-objective) updates and adaptive step-sizes; (b) enjoys any-time convergence guarantees without requiring an exact global clock; and (c) when the data is sufficiently sparse, its convergence rate for (non-)smooth, (non-)strongly-convex, and even a limited class of non-convex objectives matches the corresponding serial rate, implying a theoretical “linear speed-up”. The second algorithm, HEDGEHOG, is an asynchronous parallel version of the Exponentiated Gradient (EG) algorithm for optimization over the probability simplex (

a.k.a. Hedge in online learning), and, to our knowledge, the first asynchronous algorithm enjoying linear speed-ups under sparsity with non-SGD-style updates. Unlike previous work, ASYNCADA and HEDGEHOG and their convergence and speed-up analyses are not limited to individual coordinate-wise (i.e., “box-shaped”) constraints or smooth and strongly-convex objectives. Underlying both results is a generic analysis framework that is of independent
interest, and further applicable to distributed and delayed feedback optimization
****
Improved Precision and Recall Metric for Assessing Generative Models
Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, Timo Aila
The ability to automatically estimate the quality and coverage of the samples produced by a generative model is a vital requirement for driving algorithm research. We present an evaluation metric that can separately and reliably measure both of these aspects in image generation tasks by forming explicit, non-parametric representations of the manifolds of real and generated data. We demonstrate the effectiveness of our metric in StyleGAN and BigGAN by providing several illustrative examples where existing metrics yield uninformative or contradictory results. Furthermore, we analyze multiple design variants of StyleGAN to better understand the relationships between the model architecture, training methods, and the properties of the resulting sample distribution. In the process, we identify new variants that improve the state-of-the-art. We also perform the first principled analysis of truncation methods and identify an improved method. Finally, we extend our metric to estimate the perceptual quality of individual samples, and use this to study latent space interpolations.
****
A Direct tilde{O}(1/epsilon) Iteration Parallel Algorithm for Optimal Transport
Arun Jambulapati, Aaron Sidford, Kevin Tian
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.
****
Zero-Shot Semantic Segmentation
Maxime Bucher, Tuan-Hung VU, Matthieu Cord, Patrick Pérez
Semantic segmentation models are limited in their ability to scale to large numbers of object classes. In this paper, we introduce the new task of zero-shot semantic segmentation: learning pixel-wise classifiers for never-seen object categories with zero training examples. To this end, we present a novel architecture, ZS3Net, combining a deep visual segmentation model with an approach to generate visual representations from semantic word embeddings. By this way, ZS3Net addresses pixel classification tasks where both seen and unseen categories are faced at test time (so called generalized zero-shot classification). Performance is further improved by a self-training step that relies on automatic pseudo-labeling of pixels from unseen classes. On the two standard segmentation datasets, Pascal-VOC and Pascal-Context, we propose zero-shot benchmarks and set competitive baselines. For complex scenes as ones in the Pascal-Context dataset, we extend our approach by using a graph-context encoding to fully leverage spatial context priors coming from class-wise segmentation maps.
****
Hyperspherical Prototype Networks
Pascal Mettes, Elise van der Pol, Cees Snoek
This paper introduces hyperspherical prototype networks, which unify classification and regression with prototypes on hyperspherical output spaces. For classification, a common approach is to define prototypes as the mean output vector over training examples per class. Here, we propose to use hyperspheres as output spaces, with class prototypes defined a priori with large margin separation. We position prototypes through data-independent optimization, with an extension to incorporate priors from class semantics. By doing so, we do not require any prototype updating, we can handle any training size, and the output dimensionality is n

o longer constrained to the number of classes. Furthermore, we generalize to reg
ression, by optimizing outputs as an interpolation between two prototypes on the
 hypersphere. Since both tasks are now defined by the same loss function, they c
an be jointly trained for multi-task problems. Experimentally, we show the benef
it of hyperspherical prototype networks for classification, regression, and thei
r combination over other prototype methods, softmax cross-entropy, and mean squa
red error approaches.
****
Lower Bounds on Adversarial Robustness from Optimal Transport
Arjun Nitin Bhagoji, Daniel Cullina, Prateek Mittal
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.
****
A Nonconvex Approach for Exact and Efficient Multichannel Sparse Blind Deconvolu
tion
Qing Qu, Xiao Li, Zhihui 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Generalization of Reinforcement Learners with Working and Episodic Memory
Meire Fortunato, Melissa Tan, Ryan Faulkner, Steven Hansen, Adrià Puigdomènech B
adia, Gavin Buttimore, Charles Deck, Joel Z. Leibo, Charles Blundell
Memory is an important aspect of intelligence and plays a role in many deep rein
forcement learning models. However, little progress has been made in understandi
ng when specific memory systems help more than others and how well they generali
ze. The field also has yet to see a prevalent consistent and rigorous approach f
or evaluating agent performance on holdout data.
In this paper, we aim to develop a comprehensive methodology to test different k
inds of memory in an agent and assess how well the agent can apply what it learn
s in training to a holdout set that differs from the training set along dimensio
ns that we suggest are relevant for evaluating memory-specific generalization. T
o that end, we first construct a diverse set of memory tasks that allow us to ev
aluate test-time generalization across multiple dimensions. Second, we develop a
nd perform multiple ablations on an agent architecture that combines multiple me
mory systems, observe its baseline models, and investigate its performance again
st the task suite.
****
DTWNet: a Dynamic Time Warping Network
Xingyu Cai, Tingyang Xu, Jinfeng Yi, Junzhou Huang, Sanguthevar Rajasekaran
Dynamic Time Warping (DTW) is widely used as a similarity measure in various dom
ains. Due to its invariance against warping in the time axis, DTW provides more
meaningful discrepancy measurements between two signals than other dis- tance me
asures. In this paper, we propose a novel component in an artificial neural netw
ork. In contrast to the previous successful usage of DTW as a loss function, the
 proposed framework leverages DTW to obtain a better feature extraction. For the
 first time, the DTW loss is theoretically analyzed, and a stochastic backpropog
ation scheme is proposed to improve the accuracy and efficiency of the DTW learn
ing. We also demonstrate that the proposed framework can be used as a data analy
sis tool to perform data decomposition.
****
Learning Mean-Field Games
Xin Guo, Anran Hu, Renyuan Xu, Junzi Zhang
This paper presents a general mean-field game (GMFG) framework for simultaneous
learning and decision-making in stochastic games with a large population. It fir
st establishes the existence of a unique Nash Equilibrium to this GMFG, and expl
ains that naively combining Q-learning with the fixed-point approach in classica

l MFGs yields unstable algorithms. It then proposes a Q-learning algorithm with Boltzmann policy (GMF-Q), with analysis of convergence property and computational complexity. The experiments on repeated Ad auction problems demonstrate that this GMF-Q algorithm is efficient and robust in terms of convergence and learning accuracy. Moreover, its performance is superior in convergence, stability, and learning ability, when compared with existing algorithms for multi-agent reinforcement learning.
****
Learning Erdos-Renyi Random Graphs via Edge Detecting Queries
Zihan Li, Matthias Fresacher, Jonathan Scarlett
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.
****
Cormorant: Covariant Molecular Neural Networks
Brandon Anderson, Truong Son Hy, Risi Kondor
We propose Cormorant, a rotationally covariant neural network architecture for learning the behavior and properties of complex many-body physical systems. We apply these networks to molecular systems with two goals: learning atomic potential energy surfaces for use in Molecular Dynamics simulations, and learning ground state properties of molecules calculated by Density Functional Theory. Some of the key features of our network are that (a) each neuron explicitly corresponds to a subset of atoms; (b) the activation of each neuron is covariant to rotations, ensuring that overall the network is fully rotationally invariant. Furthermore, the non-linearity in our network is based upon tensor products and the Clebsch-Gordan decomposition, allowing the network to operate entirely in Fourier space. Cormorant significantly outperforms competing algorithms in learning molecular Potential Energy Surfaces from conformational geometries in the MD-17 dataset, and is competitive with other methods at learning geometric, energetic, electronic, and thermodynamic properties of molecules on the GDB-9 dataset.
****
Flattening a Hierarchical Clustering through Active Learning
Fabio Vitale, Anand Rajagopalan, Claudio Gentile
We investigate active learning by pairwise similarity over the leaves of trees originating from hierarchical clustering procedures. In the realizable setting, we provide a full characterization of the number of queries needed to achieve perfect reconstruction of the tree cut. In the non-realizable setting, we rely on known important-sampling procedures to obtain regret and query complexity bounds. Our algorithms come with theoretical guarantees on the statistical error and, more importantly, lend themselves to {\em linear-time} implementations in the relevant parameters of the problem. We discuss such implementations, prove running time guarantees for them, and present preliminary experiments on real-world datasets showing the compelling practical performance of our algorithms as compared to both passive learning and simple active learning baselines.
****
Random Projections and Sampling Algorithms for Clustering of High-Dimensional Polygonal Curves
Stefan Meintrup, Alexander Munteanu, 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.
****
Explicit Explore-Exploit Algorithms in Continuous State Spaces
Mikael Henaff
We present a new model-based algorithm for reinforcement learning (RL) which consists of explicit exploration and exploitation phases, and is applicable in large or
infinite state spaces. The algorithm maintains a set of dynamics models consiste

nt
with current experience and explores by finding policies which induce high dis-agreement between their state predictions. It then exploits using the refined set of models or experience gathered during exploration. We show that under realizability and optimal planning assumptions, our algorithm provably finds a near-optimal policy with a number of samples that is polynomial in a structural complexity measure which we show to be low in several natural settings. We then give a practical approximation using neural networks and demonstrate its performance and sample efficiency in practice.
****

How degenerate is the parametrization of neural networks with the ReLU activation function?

Dennis Maximilian Elbrächter, Julius Berner, Philipp Grohs

Neural network training is usually accomplished by solving a non-convex optimization problem using stochastic gradient descent. Although one optimizes over the networks parameters, the main loss function generally only depends on the realization of the neural network, i.e. the function it computes. Studying the optimization problem over the space of realizations opens up new ways to understand neural network training. In particular, usual loss functions like mean squared error and categorical cross entropy are convex on spaces of neural network realizations, which themselves are non-convex. Approximation capabilities of neural networks can be used to deal with the latter non-convexity, which allows us to establish that for sufficiently large networks local minima of a regularized optimization problem on the realization space are almost optimal. Note, however, that each realization has many different, possibly degenerate, parametrizations. In particular, a local minimum in the parametrization space needs not correspond to a local minimum in the realization space. To establish such a connection, inverse stability of the realization map is required, meaning that proximity of realizations must imply proximity of corresponding parametrizations. We present pathologies which prevent inverse stability in general, and, for shallow networks, proceed to establish a restricted space of parametrizations on which we have inverse stability w.r.t. to a Sobolev norm. Furthermore, we show that by optimizing over such restricted sets, it is still possible to learn any function which can be learned by optimization over unrestricted sets.
****

Hyperbolic Graph Convolutional Neural Networks

Ines Chami, Zhitao Ying, Christopher Ré, Jure Leskovec

Graph convolutional neural networks (GCNs) embed nodes in a graph into Euclidean space, which has been shown to incur a large distortion when embedding real-world graphs with scale-free or hierarchical structure. Hyperbolic geometry offers an exciting alternative, as it enables embeddings with much smaller distortion. However, extending GCNs to hyperbolic geometry presents several unique challenges because it is not clear how to define neural network operations, such as feature transformation and aggregation, in hyperbolic space. Furthermore, since input features are often Euclidean, it is unclear how to transform the features into hyperbolic embeddings with the right amount of curvature. Here we propose Hyperbolic Graph Convolutional Neural Network (HGCN), the first inductive hyperbolic GCN that leverages both the expressiveness of GCNs and hyperbolic geometry to learn inductive node representations for hierarchical and scale-free graphs. We derive GCNs operations in the hyperboloid model of hyperbolic space and map Euclidean input features to embeddings in hyperbolic spaces with different trainable curvature at each layer. Experiments demonstrate that HGCN learns embeddings that preserve hierarchical structure, and leads to improved performance when compared to Euclidean analogs, even with very low dimensional embeddings: compared to state-of-the-art GCNs, HGCN achieves an error reduction of up to 63.1% in ROC AUC for link prediction and of up to 47.5% in F1 score for node classification, also improving state-of-the art on the Pubmed dataset.
****

Spherical Text Embedding
Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, Jiawei Han
Unsupervised text embedding has shown great power in a wide range of NLP tasks. While text embeddings are typically learned in the Euclidean space, directional similarity is often more effective in tasks such as word similarity and document clustering, which creates a gap between the training stage and usage stage of text embedding. To close this gap, we propose a spherical generative model based on which unsupervised word and paragraph embeddings are jointly learned. To learn text embeddings in the spherical space, we develop an efficient optimization algorithm with convergence guarantee based on Riemannian optimization. Our model enjoys high efficiency and achieves state-of-the-art performances on various text embedding tasks including word similarity and document clustering.
****
Random Tessellation Forests
Shufei Ge, Shijia Wang, Yee Whye Teh, Liangliang Wang, Lloyd Elliott
Space partitioning methods such as random forests and the Mondrian process are powerful machine learning methods for multi-dimensional and relational data, and are based on recursively cutting a domain. The flexibility of these methods is often limited by the requirement that the cuts be axis aligned. The Ostomachion process and the self-consistent binary space partitioning-tree process were recently introduced as generalizations of the Mondrian process for space partitioning with non-axis aligned cuts in the plane. Motivated by the need for a multi-dimensional partitioning tree with non-axis aligned cuts, we propose the Random Tessellation Process, a framework that includes the Mondrian process as a special case. We derive a sequential Monte Carlo algorithm for inference, and provide random forest methods. Our methods are self-consistent and can relax axis-aligned constraints, allowing complex inter-dimensional dependence to be captured. We present a simulation study and analyze gene expression data of brain tissue, showing improved accuracies over other methods.
****
SpArSe: Sparse Architecture Search for CNNs on Resource-Constrained Microcontrollers
Igor Fedorov, Ryan P. Adams, Matthew Mattina, Paul Whatmough
The vast majority of processors in the world are actually microcontroller units (MCUs), which find widespread use performing simple control tasks in applications ranging from automobiles to medical devices and office equipment. The Internet of Things (IoT) promises to inject machine learning into many of these every-day objects via tiny, cheap MCUs. However, these resource-impoverished hardware platforms severely limit the complexity of machine learning models that can be deployed. For example, although convolutional neural networks (CNNs) achieve state-of-the-art results on many visual recognition tasks, CNN inference on MCUs is challenging due to severe memory limitations. To circumvent the memory challenge associated with CNNs, various alternatives have been proposed that do fit within the memory budget of an MCU, albeit at the cost of prediction accuracy. This paper challenges the idea that CNNs are not suitable for deployment on MCUs. We demonstrate that it is possible to automatically design CNNs which generalize well, while also being small enough to fit onto memory-limited MCUs. Our Sparse Architecture Search method combines neural architecture search with pruning in a single, unified approach, which learns superior models on four popular IoT datasets. The CNNs we find are more accurate and up to 7.4× smaller than previous approaches, while meeting the strict MCU working memory constraint.
****
Capacity Bounded Differential Privacy
Kamalika Chaudhuri, Jacob Imola, Ashwin Machanavajjhala
Differential privacy, a notion of algorithmic stability, is a gold standard for measuring the additional risk an algorithm's output poses to the privacy of a single record in the dataset. Differential privacy is defined as the distance between the output distribution of an algorithm on neighboring datasets that differ in one entry. In this work, we present a novel relaxation of differential

privacy, capacity bounded differential privacy, where the adversary
that distinguishes output distributions is assumed to be
capacity-bounded -- i.e. bounded not in computational power, but in
terms of the function class from which their attack algorithm is drawn. We model
adversaries in terms of restricted f-divergences between probability
distributions, and study properties of the definition and algorithms that
satisfy them.
****
Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates
Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, Daniel M. Roy
In this work, we improve upon the stepwise analysis of noisy iterative learning algorithms initiated by Pensia, Jog, and Loh (2018) and recently extended by Bu, Zou, and Veeravalli (2019). Our main contributions are significantly improved mutual information bounds for Stochastic Gradient Langevin Dynamics via data-dependent estimates. Our approach is based on the variational characterization of mutual information and the use of data-dependent priors that forecast the mini-batch gradient based on a subset of the training samples. Our approach is broadly applicable within the information-theoretic framework of Russo and Zou (2015) and Xu and Raginsky (2017). Our bound can be tied to a measure of flatness of the empirical risk surface. As compared with other bounds that depend on the squared norms of gradients, empirical investigations show that the terms in our bounds are orders of magnitude smaller.
****
Efficient Algorithms for Smooth Minimax Optimization
Kiran K. Thekumparampil, Prateek Jain, Praneeth Netrapalli, Sewoong Oh
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.
****
Uniform convergence may be unable to explain generalization in deep learning
Vaishnavh Nagarajan, 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.
****
First order expansion of convex regularized estimators
Pierre Bellec, Arun Kuchibhotla
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 exploration in linear quadratic reinforcement learning
Jack Umenberger, Mina Ferizbegovic, Thomas B. Schön, Håkan Hjalmarsson
Learning to make decisions in an uncertain and dynamic environment is a task of fundamental performance in a number of domains.
This paper concerns the problem of learning control policies for an unknown linear dynamical system so as to minimize a quadratic cost function.
We present a method, based on convex optimization, that accomplishes this task ‘robustly’, i.e., the worst-case cost, accounting for system uncertainty given the observed data, is minimized.
The method balances exploitation and exploration, exciting the system in such a way so as to reduce uncertainty in the model parameters to which the worst-case cost is most sensitive.
Numerical simulations and application to a hardware-in-the-loop servo-mechanism are used to demonstrate the approach, with appreciable performance and robustnes

s gains over alternative methods observed in both.
****
Modeling Uncertainty by Learning a Hierarchy of Deep Neural Connections
Raanan Yehezkel Rohekar, Yaniv Gurwicz, Shami Nisimov, Gal Novik
Modeling uncertainty in deep neural networks, despite recent important advances, is still an open problem. Bayesian neural networks are a powerful solution, where the prior over network weights is a design choice, often a normal distribution or other distribution encouraging sparsity. However, this prior is agnostic to the generative process of the input data, which might lead to unwarranted generalization for out-of-distribution tested data. We suggest the presence of a confounder for the relation between the input data and the discriminative function given the target label.
We propose an approach for modeling this confounder by sharing neural connectivity patterns between the generative and discriminative networks. This approach leads to a new deep architecture, where networks are sampled from the posterior of local causal structures, and coupled into a compact hierarchy. We demonstrate that sampling networks from this hierarchy, proportionally to their posterior, is efficient and enables estimating various types of uncertainties. Empirical evaluations of our method demonstrate significant improvement compared to state-of-the-art calibration and out-of-distribution detection methods.
****
Meta-Surrogate Benchmarking for Hyperparameter Optimization
Aaron Klein, Zhenwen Dai, Frank Hutter, Neil Lawrence, Javier Gonzalez
Despite the recent progress in hyperparameter optimization (HPO), available benchmarks that resemble real-world scenarios consist of a few and very large problem instances that are expensive to solve. This blocks researchers and practitioners no only from systematically running large-scale comparisons that are needed to draw statistically significant results but also from reproducing experiments that were conducted before.
This work proposes a method to alleviate these issues by means of a meta-surrogate model for HPO tasks trained on off-line generated data. The model combines a probabilistic encoder with a multi-task model such that it can generate inexpensive and realistic tasks of the class of problems of interest.
We demonstrate that benchmarking HPO methods on samples of the generative model allows us to draw more coherent and statistically significant conclusions that can be reached orders of magnitude faster than using the original tasks. We provide evidence of our findings for various HPO methods on a wide class of problems.
****
Time/Accuracy Tradeoffs for Learning a ReLU with respect to Gaussian Marginals
Surbhi Goel, Sushrut Karmalkar, Adam Klivans
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.
****
Bayesian Optimization under Heavy-tailed Payoffs
Sayak Ray Chowdhury, Aditya Gopalan
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.
****
Distribution Learning of a Random Spatial Field with a Location-Unaware Mobile Sensor
Meera Pai, Animesh Kumar
Measurement of spatial fields is of interest in environment monitoring. Recently mobile sensing has been proposed for spatial field reconstruction, which requires a smaller number of sensors when compared to the traditional paradigm of sensing with static sensors. A challenge in mobile sensing is to overcome the location uncertainty of its sensors. While GPS or other localization methods can reduc

e this uncertainty, we address a more fundamental question: can a location-unawa re mobile sensor, recording samples on a directed non-uniform random walk, learn the statistical distribution (as a function of space) of an underlying random p rocess (spatial field)? The answer is in the affirmative for Lipschitz continuou s fields, where the accuracy of our distribution-learning method increases with the number of observed field samples (sampling rate). To validate our distributi on-learning method, we have created a dataset with 43 experimental trials by mea suring sound-level along a fixed path using a location-unaware mobile sound-leve l meter.
****
State Aggregation Learning from Markov Transition Data
Yaqi Duan, Tracy Ke, Mengdi Wang
State aggregation is a popular model reduction method rooted in optimal control. It reduces the complexity of engineering systems by mapping the system’s states into a small number of meta-states. The choice of aggregation map often depends on the data analysts’ knowledge and is largely ad hoc. In this paper, we propos e a tractable algorithm that estimates the probabilistic aggregation map from th e system’s trajectory. We adopt a soft-aggregation model, where each meta-state has a signature raw state, called an anchor state. This model includes several c ommon state aggregation models as special cases. Our proposed method is a simple two- step algorithm: The first step is spectral decomposition of empirical tran sition matrix, and the second step conducts a linear transformation of singular vectors to find their approximate convex hull. It outputs the aggregation distri butions and disaggregation distributions for each meta-state in explicit forms, which are not obtainable by classical spectral methods. On the theoretical side, we prove sharp error bounds for estimating the aggregation and disaggregation d istributions and for identifying anchor states. The analysis relies on a new ent ry-wise deviation bound for singular vectors of the empirical transition matrix of a Markov process, which is of independent interest and cannot be deduced from existing literature. The application of our method to Manhattan traffic data su ccessfully generates a data-driven state aggregation map with nice interpretatio ns.
****
Reliable training and estimation of variance networks
Nicki Skafte, Martin Jørgensen, Søren Hauberg
We propose and investigate new complementary methodologies for estimating predic tive variance networks in regression neural networks. We derive a locally aware mini-batching scheme that results in sparse robust gradients, and we show how to make unbiased weight updates to a variance network. Further, we formulate a heu ristic for robustly fitting both the mean and variance networks post hoc. Finall y, we take inspiration from posterior Gaussian processes and propose a network a rchitecture with similar extrapolation properties to Gaussian processes. The pro posed methodologies are complementary, and improve upon baseline methods individ ually. Experimentally, we investigate the impact of predictive uncertainty on mu ltiple datasets and tasks ranging from regression, active learning and generativ e modeling. Experiments consistently show significant improvements in predictive uncertainty estimation over state-of-the-art methods across tasks and datasets.
****
Meta-Learning with Implicit Gradients
Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, Sergey Levine
A core capability of intelligent systems is the ability to quickly learn new tas ks by drawing on prior experience. Gradient (or optimization) based meta-learnin g has recently emerged as an effective approach for few-shot learning. In this f ormulation, meta-parameters are learned in the outer loop, while task-specific m odels are learned in the inner-loop, by using only a small amount of data from t he current task. A key challenge in scaling these approaches is the need to diff erentiate through the inner loop learning process, which can impose considerable computational and memory burdens. By drawing upon implicit differentiation, we develop the implicit MAML algorithm, which depends only on the solution to the i nner level optimization and not the path taken by the inner loop optimizer. This

effectively decouples the meta-gradient computation from the choice of inner loop optimizer. As a result, our approach is agnostic to the choice of inner loop optimizer and can gracefully handle many gradient steps without vanishing gradients or memory constraints. Theoretically, we prove that implicit MAML can compute accurate meta-gradients with a memory footprint that is, up to small constant factors, no more than that which is required to compute a single inner loop gradient and at no overall increase in the total computational cost. Experimentally, we show that these benefits of implicit MAML translate into empirical gains on few-shot image recognition benchmarks.
****
Differentially Private Markov Chain Monte Carlo
Mikko Heikkilä, Joonas Jälkö, Onur Dikmen, Antti Honkela
Recent developments in differentially private (DP) machine learning and DP Bayesian learning have enabled learning under strong privacy guarantees for the training data subjects. In this paper, we further extend the applicability of DP Bayesian learning by presenting the first general DP Markov chain Monte Carlo (MCMC) algorithm whose privacy-guarantees are not subject to unrealistic assumptions on Markov chain convergence and that is applicable to posterior inference in arbitrary models. Our algorithm is based on a decomposition of the Barker acceptance test that allows evaluating the Rényi DP privacy cost of the accept-reject choice. We further show how to improve the DP guarantee through data subsampling and approximate acceptance tests.
****
Universal Boosting Variational Inference
Trevor Campbell, Xinglong Li
Boosting variational inference (BVI) approximates an intractable probability density by iteratively building up a mixture of simple component distributions one at a time, using techniques from sparse convex optimization to provide both computational scalability and approximation error guarantees. But the guarantees have strong conditions that do not often hold in practice, resulting in degenerate component optimization problems; and we show that the ad-hoc regularization used to prevent degeneracy in practice can cause BVI to fail in unintuitive ways. We thus develop universal boosting variational inference (UBVI), a BVI scheme that exploits the simple geometry of probability densities under the Hellinger metric to prevent the degeneracy of other gradient-based BVI methods, avoid difficult joint optimizations of both component and weight, and simplify fully-corrective weight optimizations. We show that for any target density and any mixture component family, the output of UBVI converges to the best possible approximation in the mixture family, even when the mixture family is misspecified. We develop a scalable implementation based on exponential family mixture components and standard stochastic optimization techniques. Finally, we discuss statistical benefits of the Hellinger distance as a variational objective through bounds on posterior probability, moment, and importance sampling errors. Experiments on multiple datasets and models show that UBVI provides reliable, accurate posterior approximations.
****
LIIR: Learning Individual Intrinsic Reward in Multi-Agent Reinforcement Learning
Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, Dacheng Tao
A great challenge in cooperative decentralized multi-agent reinforcement learning (MARL) is generating diversified behaviors for each individual agent when receiving only a team reward. Prior studies have paid much effort on reward shaping or designing a centralized critic that can discriminatively credit the agents. In this paper, we propose to merge the two directions and learn each agent an intrinsic reward function which diversely stimulates the agents at each time step. Specifically, the intrinsic reward for a specific agent will be involved in computing a distinct proxy critic for the agent to direct the updating of its individual policy. Meanwhile, the parameterized intrinsic reward function will be updated towards maximizing the expected accumulated team reward from the environment so that the objective is consistent with the original MARL problem. The proposed method is referred to as learning individual intrinsic reward (LIIR) in MARL.

We compare LIIR with a number of state-of-the-art MARL methods on battle games in StarCraft II. The results demonstrate the effectiveness of LIIR, and we show LIIR can assign each individual agent an insightful intrinsic reward per time step.
****
A Normative Theory for Causal Inference and Bayes Factor Computation in Neural Circuits
Wenhao Zhang, Si Wu, Brent Doiron, Tai Sing Lee
This study provides a normative theory for how Bayesian causal inference can be implemented in neural circuits. In both cognitive processes such as causal reasoning and perceptual inference such as cue integration, the nervous systems need to choose different models representing the underlying causal structures when making inferences on external stimuli. In multisensory processing, for example, the nervous system has to choose whether to integrate or segregate inputs from different sensory modalities to infer the sensory stimuli, based on whether the inputs are from the same or different sources. Making this choice is a model selection problem requiring the computation of Bayes factor, the ratio of likelihoods between the integration and the segregation models. In this paper, we consider the causal inference in multisensory processing and propose a novel generative model based on neural population code that takes into account both stimulus feature and stimulus reliability in the inference. In the case of circular variables such as heading direction, our normative theory yields an analytical solution for computing the Bayes factor, with a clear geometric interpretation, which can be implemented by simple additive mechanisms with neural population code. Numerical simulation shows that the tunings of the neurons computing Bayes factor are consistent with the "opposite neurons" discovered in dorsal medial superior temporal (MSTd) and the ventral intraparietal (VIP) areas for visual-vestibular processing. This study illuminates a potential neural mechanism for causal inference in the brain.
****
The Geometry of Deep Networks: Power Diagram Subdivision
Randall Balestriero, Romain Cosentino, Behnaam Aazhang, Richard Baraniuk
We study the geometry of deep (neural) networks (DNs) with piecewise affine and convex nonlinearities.
The layers of such DNs have been shown to be max-affine spline operators (MASOs) that partition their input space and apply a region-dependent affine mapping to their input to produce their output.
We demonstrate that each MASO layer's input space partitioning corresponds to a power diagram (an extension of the classical Voronoi tiling) with a number of regions that grows exponentially with respect to the number of units (neurons).
We further show that a composition of MASO layers (e.g., the entire DN) produces a progressively subdivided power diagram and provide its analytical form.
The subdivision process constrains the affine maps on the potentially exponentially many power diagram regions with respect to the number of neurons to greatly reduce their complexity.
For classification problems, we obtain a formula for a MASO DN's decision boundary in the input space plus a measure of its curvature that depends on the DN's nonlinearities, weights, and architecture.
Numerous numerical experiments support and extend our theoretical results.
****
Visual Sequence Learning in Hierarchical Prediction Networks and Primate Visual Cortex

In this paper we developed a computational hierarchical network model to understand the spatiotemporal sequence learning effects observed in the primate visual cortex. The model is a hierarchical recurrent neural model that learns to predict video sequences using the incoming video signals as teaching signals.
The model performs fast feedforward analysis using a deep convolutional neural network with sparse convolution and feedback synthesis using a stack of LSTM modules. The network learns a representational hierarchy by minimizing its predict

ion errors of the incoming signals at each level of the hierarchy. We found that recurrent feedback in this network lead to the development of semantic cluster of global movement patterns in the population codes of the units at the lower levels of the hierarchy. These representations facilitate the learning of relationship among movement patterns, yielding state-of-the-art performance in long range video sequence predictions on benchmark datasets. Without further tuning, this model automatically exhibits the neurophysiological correlates of visual sequence memories that we observed in the early visual cortex of awake monkeys, suggesting the principle of self-supervised prediction learning might be relevant to understanding the cortical mechanisms of representational learning.
****
Equal Opportunity in Online Classification with Partial Feedback
Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, Steven Z. Wu
We study an online classification problem with partial feedback in which individuals arrive one at a time from a fixed but unknown distribution, and must be classified as positive or negative. Our algorithm only observes the true label of an individual if they are given a positive classification. This setting captures many classification problems for which fairness is a concern: for example, in criminal recidivism prediction, recidivism is only observed if the inmate is released; in lending applications, loan repayment is only observed if the loan is granted. We require that our algorithms satisfy common statistical fairness constraints (such as equalizing false positive or negative rates --- introduced as "equal opportunity" in Hardt et al. (2016)) at every round, with respect to the underlying distribution. We give upper and lower bounds characterizing the cost of this constraint in terms of the regret rate (and show that it is mild), and give an oracle efficient algorithm that achieves the upper bound.
****
Semi-Parametric Efficient Policy Learning with Continuous Actions
Victor Chernozhukov, Mert Demirer, Greg Lewis, Vasilis Syrgkanis
We consider off-policy evaluation and optimization with continuous action spaces. We focus on observational data where the data collection policy is unknown and needs to be estimated from data. We take a semi-parametric approach where the value function takes a known parametric form in the treatment, but we are agnostic on how it depends on the observed contexts. We propose a doubly robust off-policy estimate for this setting and show that off-policy optimization based on this doubly robust estimate is robust to estimation errors of the policy function or the regression model. We also show that the variance of our off-policy estimate achieves the semi-parametric efficiency bound. Our results also apply if the model does not satisfy our semi-parametric form but rather we measure regret in terms of the best projection of the true value function to this functional space. Our work extends prior approaches of policy optimization from observational data that only considered discrete actions. We provide an experimental evaluation of our method in a synthetic data example motivated by optimal personalized pricing.
****
Concentration of risk measures: A Wasserstein distance approach
Sanjay P. Bhat, Prashanth L.A.
Known finite-sample concentration bounds for the Wasserstein distance between the empirical and true distribution of a random variable are used to derive a two-sided concentration bound for the error between the true conditional value-at-risk (CVaR) of a (possibly unbounded) random variable and a standard estimate of its CVaR computed from an i.i.d. sample. The bound applies under fairly general assumptions on the random variable, and improves upon previous bounds which were either one sided, or applied only to bounded random variables. Specializations of the bound to sub-Gaussian and sub-exponential random variables are also derived. A similar procedure is followed to derive concentration bounds for the error between the true and estimated Cumulative Prospect Theory (CPT) value of a random variable, in cases where the random variable is bounded or sub-Gaussian. These bounds are shown to match a known bound in the bounded case, and improve upon the known bound in the sub-Gaussian case. The usefulness of the bounds is illu

strated through an algorithm, and corresponding regret bound for a stochastic bandit problem, where the underlying risk measure to be optimized is CVaR.
****

Interior-Point Methods Strike Back: Solving the Wasserstein Barycenter Problem
DongDong Ge, Haoyue Wang, Zikai Xiong, Yinyu Ye
Computing the Wasserstein barycenter of a set of probability measures under the optimal transport metric can quickly become prohibitive for traditional second-order algorithms, such as interior-point methods, as the support size of the measures increases. In this paper, we overcome the difficulty by developing a new adapted interior-point method that fully exploits the problem's special matrix structure to reduce the iteration complexity and speed up the Newton procedure. Different from regularization approaches, our method achieves a well-balanced tradeoff between accuracy and speed. A numerical comparison on various distributions with existing algorithms exhibits the computational advantages of our approach. Moreover, we demonstrate the practicality of our algorithm on image benchmark problems including MNIST and Fashion-MNIST.
****

Coda: An End-to-End Neural Program Decompiler
Cheng Fu, Huili Chen, Haolan Liu, Xinyun Chen, Yuandong Tian, Farinaz Koushanfar, Jishen Zhao
Reverse engineering of binary executables is a critical problem in the computer security domain. On the one hand, malicious parties may recover interpretable source codes from the software products to gain commercial advantages. On the other hand, binary decompilation can be leveraged for code vulnerability analysis and malware detection. However, efficient binary decompilation is challenging. Conventional decompilers have the following major limitations: (i) they are only applicable to specific source-target language pair, hence incurs undesired development cost for new language tasks; (ii) their output high-level code cannot effectively preserve the correct functionality of the input binary; (iii) their output program does not capture the semantics of the input and the reversed program is hard to interpret. To address the above problems, we propose Coda1, the first end-to-end neural-based framework for code decompilation. Coda decomposes the decompilation task into of two key phases: First, Coda employs an instruction type-aware encoder and a tree decoder for generating an abstract syntax tree (AST) with attention feeding during the code sketch generation stage. Second, Coda then updates the code sketch using an iterative error correction machine guided by an ensembled neural error predictor. By finding a good approximate candidate and then fixing it towards perfect, Coda achieves superior with performance compared to baseline approaches. We assess Coda’s performance with extensive experiments on various benchmarks. Evaluation results show that Coda achieves an average of 82% program recovery accuracy on unseen binary samples, where the state-of-the-art decompilers yield 0% accuracy. Furthermore, Coda outperforms the sequence-to-sequence model with attention by a margin of 70% program accuracy. Our work reveals the vulnerability of binary executables and imposes a new threat to the protection of Intellectual Property (IP) for software development.
****

GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism
Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, zhifeng Chen
Scaling up deep neural network capacity has been known as an effective approach to improving model quality for several different machine learning tasks. In many cases, increasing model capacity beyond the memory limit of a single accelerator has required developing special algorithms or infrastructure. These solutions are often architecture-specific and do not transfer to other machine learning tasks. To address the need for efficient and task-independent model parallelism, we introduce TensorPipe, a pipeline parallelism library that allows scaling any network that can be expressed as a sequence of layers. By pipelining different sub-sequences of layers on separate accelerators, TensorPipe provides the flexibility of scaling a variety of different networks to gigantic sizes efficiently. Moreover, TensorPipe utilizes a novel batch-splitting pipelining algorithm, resu

lting in almost linear speedup when a model is partitioned across multiple accel erators. We demonstrate the advantages of TensorPipe by training large-sca le neural networks on two different tasks with distinct network architecture s: (i)Image Classification: We train a 557-million-parameter AmoebaNet model and attain a top-1 accuracy of 84.4% on ImageNet-2012, (ii)Multilingual Neural Mach ine Translation: We train a single 6-billion-parameter, 128-layer Transformer mo del on a corpus spanning over 100 languages and achieve better quality than all bilingual models.
****
DiskANN: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node
Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Kris hnawamy, Rohan Kadekodi
Current state-of-the-art approximate nearest neighbor search (ANNS) algorithms generate indices that must be stored in main memory for fast high-recall search. This makes them expensive and limits the size of the dataset. We present a new graph-based indexing and search system called DiskANN that can index, store, and search a billion point database on a single workstation with just 64G B
RAM and an inexpensive solid-state drive (SSD). Contrary to current wisdom, we demonstrate that the SSD-based indices built by DiskANN can meet all three desiderata for large-scale ANNS: high-recall, low query latency and high density (points indexed per node). On the billion point SIFT1B bigann dataset, DiskANN serves > 5000 queries a second with < 3ms mean latency and 95%+ 1-recall@1 on a 16 core machine, where state-of-the-art billion-point ANNS algorithms with similar memory footprint like FAISS and IVFOADC+G+P plateau at around 50% 1-recall@1. Alternately, in the high recall regime, DiskANN can index and serve 5 – 10x more points per node compared to state-of-the-art graph-based methods such as HNSW and NSG. Finally, as part of our overall DiskANN system, we introduce Vamana, a new graph-based ANNS index that is more versatile than the graph indices even for in-memory indices.
****
Linear Stochastic Bandits Under Safety Constraints
Sanae Amani, Mahnoosh Alizadeh, Christos Thrampoulidis
Bandit algorithms have various application in safety-critical systems, where it is important to respect the system constraints that rely on the bandit's unknown parameters at every round. In this paper, we formulate a linear stochastic mult i-armed bandit problem with safety constraints that depend (linearly) on an unkn own parameter vector. As such, the learner is unable to identify all safe action s and must act conservatively in ensuring that her actions satisfy the safety co nstraint at all rounds (at least with high probability). For these bandits, we p ropose a new UCB-based algorithm called Safe-LUCB, which includes necessary modi fications to respect safety constraints. The algorithm has two phases. During th e pure exploration phase the learner chooses her actions at random from a restri cted set of safe actions with the goal of learning a good approximation of the e ntire unknown safe set. Once this goal is achieved, the algorithm begins a safe exploration-exploitation phase where the learner gradually expands their estimat e of the set of safe actions while controlling the growth of regret. We provide a general regret bound for the algorithm, as well as a problem dependent bound t hat is connected to the location of the optimal action within the safe set. We t hen propose a modified heuristic that exploits our problem dependent analysis to improve the regret.
****
Power analysis of knockoff filters for correlated designs
Jingbo Liu, Philippe Rigollet
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.
****
Implicitly learning to reason in first-order logic

Vaishak Belle, Brendan Juba
We consider the problem of answering queries about formulas of first-order logic based on background knowledge partially represented explicitly as other formulas, and partially represented as examples independently drawn from a fixed probability distribution. PAC semantics, introduced by Valiant, is one rigorous, general proposal for learning to reason in formal languages: although weaker than classical entailment, it allows for a powerful model theoretic framework for answering queries while requiring minimal assumptions about the form of the distribution in question. To date, however, the most significant limitation of that approach, and more generally most machine learning approaches with robustness guarantees, is that the logical language is ultimately essentially propositional, with finitely many atoms. Indeed, the theoretical findings on the learning of relational theories in such generality have been resoundingly negative. This is despite the fact that first-order logic is widely argued to be most appropriate for representing human knowledge.
In this work, we present a new theoretical approach to robustly learning to reason in first-order logic, and consider universally quantified clauses over a countably infinite domain. Our results exploit symmetries exhibited by constants in the language, and generalize the notion of implicit learnability to show how queries can be computed against (implicitly) learned first-order background knowledge.
****
Low-Rank Bandit Methods for High-Dimensional Dynamic Pricing
Jonas W. Mueller, Vasilis Syrgkanis, Matt Taddy
We consider dynamic pricing with many products under an evolving but low-dimensional demand model. Assuming the temporal variation in cross-elasticities exhibits low-rank structure based on fixed (latent) features of the products, we show that the revenue maximization problem reduces to an online bandit convex optimization with side information given by the observed demands. We design dynamic pricing algorithms whose revenue approaches that of the best fixed price vector in hindsight, at a rate that only depends on the intrinsic rank of the demand model and not the number of products. Our approach applies a bandit convex optimization algorithm in a projected low-dimensional space spanned by the latent product features, while simultaneously learning this span via online singular value decomposition of a carefully-crafted matrix containing the observed demands.
****
Learning Stable Deep Dynamics Models
J. Zico Kolter, Gaurav Manek
Deep networks are commonly used to model dynamical systems, predicting how the state of a system will evolve over time (either autonomously or in response to control inputs). Despite the predictive power of these systems, it has been difficult to make formal claims about the basic properties of the learned systems. In this paper, we propose an approach for learning dynamical systems that are guaranteed to be stable over the entire state space. The approach works by jointly learning a dynamics model and Lyapunov function that guarantees non-expansiveness of the dynamics under the learned Lyapunov function. We show that such learning systems are able to model simple dynamical systems and can be combined with additional deep generative models to learn complex dynamics, such as video textures, in a fully end-to-end fashion.
****
Beyond the Single Neuron Convex Barrier for Neural Network Certification
Gagandeep Singh, Rupanshu Ganvir, Markus Püschel, Martin Vechev
We propose a new parametric framework, called k-ReLU, for computing precise
and scalable convex relaxations used to certify neural networks. The key idea is to
approximate the output of multiple ReLUs in a layer jointly instead of separately.
This joint relaxation captures dependencies between the inputs to different ReLUs
in a layer and thus overcomes the convex barrier imposed by the single neuron

triangle relaxation and its approximations. The framework is parametric in the
number of k ReLUs it considers jointly and can be combined with existing verifie
rs
in order to improve their precision. Our experimental results show that k-ReLU e
n-
ables significantly more precise certification than existing state-of-the-art ve
rifiers
while maintaining scalability.
****
Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Mode
ls
Yuge Shi, Siddharth N, Brooks Paige, Philip Torr
Learning generative models that span multiple data modalities, such as vision an
d language, is often motivated by the desire to learn more useful, generalisable
 representations that faithfully capture common underlying factors between the m
odalities. In this work, we characterise successful learning of such models as t
he fulfilment of four criteria: i) implicit latent decomposition into shared and
 private subspaces, ii) coherent joint generation over all modalities, iii) cohe
rent cross-generation across individual modalities, and iv) improved model learn
ing for individual modalities through multi-modal integration. Here, we propose
a mixture-of-experts multi-modal variational autoencoder (MMVAE) for learning of
 generative models on different sets of modalities, including a challenging imag
e <-> language dataset, and demonstrate its ability to satisfy all four criteria
, both qualitatively and quantitatively.
****
Language as an Abstraction for Hierarchical Deep Reinforcement Learning
YiDing Jiang, Shixiang (Shane) Gu, Kevin P. Murphy, Chelsea Finn
Solving complex, temporally-extended tasks is a long-standing problem in reinfor
cement learning (RL). We hypothesize that one critical element of solving such p
roblems is the notion of compositionality. With the ability to learn sub-skills
that can be composed to solve longer tasks, i.e. hierarchical RL, we can acquire
 temporally-extended behaviors. However, acquiring effective yet general abstrac
tions for hierarchical RL is remarkably challenging. In this paper, we propose t
o use language as the abstraction, as it provides unique compositional structure
, enabling fast learning and combinatorial generalization, while retaining treme
ndous flexibility, making it suitable for a variety of problems. Our approach le
arns an instruction-following low-level policy and a high-level policy that can
reuse abstractions across tasks, in essence, permitting agents to reason using s
tructured language. To study compositional task learning, we introduce an open-s
ource object interaction environment built using the MuJoCo physics engine and t
he CLEVR engine. We find that, using our approach, agents can learn to solve to
diverse, temporally-extended tasks such as object sorting and multi-object rearr
angement, including from raw pixel observations. Our analysis find that the comp
ositional nature of language is critical for learning and systematically general
izing sub-skills in comparison to non-compositional abstractions that use the sa
me supervision.
****
High-dimensional multivariate forecasting with low-rank Gaussian Copula Processe
s
David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, Jan Gas
thaus
Predicting the dependencies between observations from multiple time series is cr
itical for applications such as anomaly detection, financial risk management, ca
usal analysis, or demand forecasting.
However, the computational and numerical difficulties of estimating time-varying
 and high-dimensional covariance matrices often limits existing methods to handl
ing at most a few hundred dimensions or requires making strong assumptions on th
e dependence between series.
We propose to combine an RNN-based time series model with a Gaussian copula proc
ess output model with a low-rank covariance structure to reduce the computationa

l complexity and handle non-Gaussian marginal distributions.
This permits to drastically reduce the number of parameters and consequently allows the modeling of time-varying correlations of thousands of time series. We show on several real-world datasets that our method provides significant accuracy improvements over state-of-the-art baselines and perform an ablation study analyzing the contributions of the different components of our model.
****
Learning Macroscopic Brain Connectomes via Group-Sparse Factorization
Farzane Aminmansour, Andrew Patterson, Lei Le, Yisu Peng, Daniel Mitchell, Franco Pestilli, Cesar F. Caiafa, Russell Greiner, Martha White
Mapping structural brain connectomes for living human brains typically requires expert analysis and rule-based models on diffusion-weighted magnetic resonance imaging. A data-driven approach, however, could overcome limitations in such rule-based approaches and improve precision mappings for individuals. In this work, we explore a framework that facilitates applying learning algorithms to automatically extract brain connectomes. Using a tensor encoding, we design an objective with a group-regularizer that prefers biologically plausible fascicle structure. We show that the objective is convex and has unique solutions, ensuring identifiable connectomes for an individual. We develop an efficient optimization strategy for this extremely high-dimensional sparse problem, by reducing the number of parameters using a greedy algorithm designed specifically for the problem. We show that this greedy algorithm significantly improves on a standard greedy algorithm, called Orthogonal Matching Pursuit. We conclude with an analysis of the solutions found by our method, showing we can accurately reconstruct the diffusion information while maintaining contiguous fascicles with smooth direction changes.
****
Optimal Sketching for Kronecker Product Regression and Low Rank Approximation
Huaian Diao, Rajesh Jayaram, Zhao Song, Wen Sun, 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.
****
Deep Gamblers: Learning to Abstain with Portfolio Theory
Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R. Salakhutdinov, Louis-Philippe Morency, Masahito Ueda
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.
****
DRUM: End-To-End Differentiable Rule Mining On Knowledge Graphs
Ali Sadeghian, Mohammadreza Armandpour, Patrick Ding, Daisy Zhe Wang
In this paper, we study the problem of learning probabilistic logical rules for inductive and interpretable link prediction. Despite the importance of inductive link prediction, most previous works focused on transductive link prediction and cannot manage previously unseen entities. Moreover, they are black-box models that are not easily explainable for humans. We propose DRUM, a scalable and differentiable approach for mining first-order logical rules from knowledge graphs that resolves these problems. We motivate our method by making a connection between learning confidence scores for each rule and low-rank tensor approximation. DRUM uses bidirectional RNNs to share useful information across the tasks of learning rules for different relations. We also empirically demonstrate the efficiency of DRUM over existing rule mining methods for inductive link prediction on a variety of benchmark datasets.
****
Combinatorial Inference against Label Noise
Paul Hongsuck Seo, Geeho Kim, Bohyung 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-auth ors prior to requesting a name change in the electronic proceedings.
****

Localized Structured Prediction

Carlo Ciliberto, Francis Bach, Alessandro Rudi

Key to structured prediction is exploiting the problem's structure to simplify t he learning process. A major challenge arises when data exhibit a local structur e (i.e., are made ``by parts'') that can be leveraged to better approximate the relation between (parts of) the input and (parts of) the output. Recent literatu re on signal processing, and in particular computer vision, shows that capturing these aspects is indeed essential to achieve state-of-the-art performance. Howe ver, in this context algorithms are typically derived on a case-by-case basis. I n this work we propose the first theoretical framework to deal with part-based d ata from a general perspective and study a novel method within the setting of st atistical learning theory. Our analysis is novel in that it explicitly quantifie s the benefits of leveraging the part-based structure of a problem on the learni ng rates of the proposed estimator.
****

Fast Low-rank Metric Learning for Large-scale and High-dimensional Data

Han Liu, Zhizhong Han, Yu-Shen Liu, Ming Gu

Low-rank metric learning aims to learn better discrimination of data subject to low-rank constraints. It keeps the intrinsic low-rank structure of datasets and reduces the time cost and memory usage in metric learning. However, it is still a challenge for current methods to handle datasets with both high dimensions and large numbers of samples. To address this issue, we present a novel fast low-ra nk metric learning (FLRML) method. FLRML casts the low-rank metric learning prob lem into an unconstrained optimization on the Stiefel manifold, which can be eff iciently solved by searching along the descent curves of the manifold. FLRML sig nificantly reduces the complexity and memory usage in optimization, which makes the method scalable to both high dimensions and large numbers of samples. Furthe rmore, we introduce a mini-batch version of FLRML to make the method scalable to larger datasets which are hard to be loaded and decomposed in limited memory. T he outperforming experimental results show that our method is with high accuracy and much faster than the state-of-the-art methods under several benchmarks with large numbers of high-dimensional data. Code has been made available at https:/ /github.com/highan911/FLRML.
****

Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, Jeffrey Pennington

A longstanding goal in deep learning research has been to precisely characterize training and generalization. However, the often complex loss landscapes of neur al networks have made a theory of learning dynamics elusive. In this work, we sh ow that for wide neural networks the learning dynamics simplify considerably and that, in the infinite width limit, they are governed by a linear model obtained from the first-order Taylor expansion of the network around its initial paramet ers. Furthermore, mirroring the correspondence between wide Bayesian neural netw orks and Gaussian processes, gradient-based training of wide neural networks wit h a squared loss produces test set predictions drawn from a Gaussian process wit h a particular compositional kernel. While these theoretical results are only ex act in the infinite width limit, we nevertheless find excellent empirical agreem ent between the predictions of the original network and those of the linearized version even for finite practically-sized networks. This agreement is robust acr oss different architectures, optimization methods, and loss functions.
****

Retrosynthesis Prediction with Conditional Graph Logic Network

Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, Le Song

Retrosynthesis is one of the fundamental problems in organic chemistry. The task is to identify reactants that can be used to synthesize a specified product mol

ecule. Recently, computer-aided retrosynthesis is finding renewed interest from both chemistry and computer science communities. Most existing approaches rely on template-based models that define subgraph matching rules, but whether or not a chemical reaction can proceed is not defined by hard decision rules. In this work, we propose a new approach to this task using the Conditional Graph Logic Network, a conditional graphical model built upon graph neural networks that learns when rules from reaction templates should be applied, implicitly considering whether the resulting reaction would be both chemically feasible and strategic. We also propose an efficient hierarchical sampling to alleviate the computation cost. While achieving a significant improvement of 8.2% over current state-of-the-art methods on the benchmark dataset, our model also offers interpretations for the prediction.
****
Efficient Pure Exploration in Adaptive Round model
Tianyuan Jin, Jieming SHI, Xiaokui Xiao, Enhong 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.
****
Unsupervised Emergence of Egocentric Spatial Structure from Sensorimotor Prediction
Alban Laflaquière, Michael Garcia Ortiz
Despite its omnipresence in robotics application, the nature of spatial knowledge and the mechanisms that underlie its emergence in autonomous agents are still poorly understood. Recent theoretical works suggest that the Euclidean structure of space induces invariants in an agent’s raw sensorimotor experience. We hypothesize that capturing these invariants is beneficial for sensorimotor prediction and that, under certain exploratory conditions, a motor representation capturing the structure of the external space should emerge as a byproduct of learning to predict future sensory experiences. We propose a simple sensorimotor predictive scheme, apply it to different agents and types of exploration, and evaluate the pertinence of these hypotheses. We show that a naive agent can capture the topology and metric regularity of its sensor’s position in an egocentric spatial frame without any a priori knowledge, nor extraneous supervision.
****
Adversarial Robustness through Local Linearization
Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, Pushmeet Kohli
Adversarial training is an effective methodology for training deep neural networks that are robust against adversarial, norm-bounded perturbations. However, the computational cost of adversarial training grows prohibitively as the size of the model and number of input dimensions increase. Further, training against less expensive and therefore weaker adversaries produces models that are robust against weak attacks but break down under attacks that are stronger. This is often attributed to the phenomenon of gradient obfuscation; such models have a highly non-linear loss surface in the vicinity of training examples, making it hard for gradient-based attacks to succeed even though adversarial examples still exist. In this work, we introduce a novel regularizer that encourages the loss to behave linearly in the vicinity of the training data, thereby penalizing gradient obfuscation while encouraging robustness. We show via extensive experiments on CIFAR-10 and ImageNet, that models trained with our regularizer avoid gradient obfuscation and can be trained significantly faster than adversarial training. Using this regularizer, we exceed current state of the art and achieve 47% adversarial accuracy for ImageNet with L-infinity norm adversarial perturbations of radius 4/255 under an untargeted, strong, white-box attack. Additionally, we match state of the art results for CIFAR-10 at 8/255.
****
Generalized Off-Policy Actor-Critic
Shangtong Zhang, Wendelin Boehmer, Shimon Whiteson

We propose a new objective, the counterfactual objective, unifying existing obje ctives for off-policy policy gradient algorithms in the continuing reinforcement learning (RL) setting. Compared to the commonly used excursion objective, which can be misleading about the performance of the target policy when deployed, our new objective better predicts such performance. We prove the Generalized Off-Po licy Policy Gradient Theorem to compute the policy gradient of the counterfactua l objective and use an emphatic approach to get an unbiased sample from this pol icy gradient, yielding the Generalized Off-Policy Actor-Critic (Geoff-PAC) algor ithm. We demonstrate the merits of Geoff-PAC over existing algorithms in Mujoco robot simulation tasks, the first empirical success of emphatic algorithms in pr evailing deep RL benchmarks.
****
Average Individual Fairness: Algorithms, Generalization and Experiments
Saeed Sharifi-Malvajerdi, Michael Kearns, Aaron Roth
We propose a new family of fairness definitions for classification problems that combine some of the best properties of both statistical and individual notions of fairness. We posit not only a distribution over individuals, but also a distr ibution over (or collection of) classification tasks. We then ask that standard statistics (such as error or false positive/negative rates) be (approximately) e qualized across individuals, where the rate is defined as an expectation over th e classification tasks. Because we are no longer averaging over coarse groups (s uch as race or gender), this is a semantically meaningful individual-level const raint. Given a sample of individuals and problems, we design an oracle-efficient algorithm (i.e. one that is given access to any standard, fairness-free learnin g heuristic) for the fair empirical risk minimization task. We also show that gi ven sufficiently many samples, the ERM solution generalizes in two directions: b oth to new individuals, and to new classification tasks, drawn from their corres ponding distributions. Finally we implement our algorithm and empirically verify its effectiveness.
****
Comparing distributions: ℓ_1 geometry improves kernel two-sample testing
meyer scetbon, Gael Varoquaux
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.
****
Nonstochastic Multiarmed Bandits with Unrestricted Delays
Tobias Sommer Thune, Nicolò Cesa-Bianchi, 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Approximate Bayesian Inference for a Mechanistic Model of Vesicle Release at a R ibbon Synapse
Cornelius Schröder, Ben James, Leon Lagnado, Philipp Berens
The inherent noise of neural systems makes it difficult to construct models whic h accurately capture experimental measurements of their activity. While much res earch has been done on how to efficiently model neural activity with descriptive models such as linear-nonlinear-models (LN), Bayesian inference for mechanistic models has received considerably less attention. One reason for this is that th ese models typically lead to intractable likelihoods and thus make parameter inf erence difficult. Here, we develop an approximate Bayesian inference scheme for a fully stochastic, biophysically inspired model of glutamate release at the ri bbon synapse, a highly specialized synapse found in different sensory systems. T he model translates known structural features of the ribbon synapse into a set o f stochastically coupled equations. We approximate the posterior distributions b y updating a parametric prior distribution via Bayesian updating rules and show that model parameters can be efficiently estimated for synthetic and experimenta

l data from in vivo two-photon experiments in the zebrafish retina. Also, we find that the model captures complex properties of the synaptic release such as the temporal precision and outperforms a standard GLM. Our framework provides a viable path forward for linking mechanistic models of neural activity to measured data.
****
Data-dependent Sample Complexity of Deep Neural Networks via Lipschitz Augmentation
Colin Wei, Tengyu Ma
Existing Rademacher complexity bounds for neural networks rely only on norm control of the weight matrices and depend exponentially on depth via a product of the matrix norms. Lower bounds show that this exponential dependence on depth is unavoidable when no additional properties of the training data are considered. We suspect that this conundrum comes from the fact that these bounds depend on the training data only through the margin. In practice, many data-dependent techniques such as Batchnorm improve the generalization performance. For feedforward neural nets as well as RNNs, we obtain tighter Rademacher complexity bounds by considering additional data-dependent properties of the network: the norms of the hidden layers of the network, and the norms of the Jacobians of each layer with respect to all previous layers. Our bounds scale polynomially in depth when these empirical quantities are small, as is usually the case in practice. To obtain these bounds, we develop general tools for augmenting a sequence of functions to make their composition Lipschitz and then covering the augmented functions. Inspired by our theory, we directly regularize the network's Jacobians during training and empirically demonstrate that this improves test performance.
****
Semi-supervisedly Co-embedding Attributed Networks
Zaiqiao Meng, Shangsong Liang, Jinyuan Fang, Teng Xiao
Deep generative models (DGMs) have achieved remarkable advances. Semi-supervised variational auto-encoders (SVAE) as a classical DGM offers a principled framework to effective generalize from small labelled data to large unlabelled ones, but it is difficult to incorporate rich unstructured relationships within the multiple heterogeneous entities. In this paper, to deal with the problem, we present a semi-supervised co-embedding model for attributed networks (SCAN) based on the generalized SVAE for the heterogeneous data, which collaboratively learns low-dimensional vector representations of both nodes and attributes for partially labelled attributed networks semi-supervisedly. The node and attribute embeddings obtained in a unified manner by our SCAN can benefit not only for capturing the proximities between nodes but also the affinities between nodes and attributes. Moreover, our model also trains a discriminative network to learn the label predictive distribution of nodes. Experimental results on real-world networks demonstrate that our model yields excellent performance in a number of applications such as attribute inference, user profiling and node classification compared to the state-of-the-art baselines.
****
Adaptive Auxiliary Task Weighting for Reinforcement Learning
Xingyu Lin, Harjatin Baweja, George Kantor, David Held
Reinforcement learning is known to be sample inefficient, preventing its application to many real-world problems, especially with high dimensional observations like images. Transferring knowledge from other auxiliary tasks is a powerful tool for improving the learning efficiency. However, the usage of auxiliary tasks has been limited so far due to the difficulty in selecting and combining different auxiliary tasks. In this work, we propose a principled online learning algorithm that dynamically combines different auxiliary tasks to speed up training for reinforcement learning. Our method is based on the idea that auxiliary tasks should provide gradient directions that, in the long term, help to decrease the loss of the main task. We show in various environments that our algorithm can effectively combine a variety of different auxiliary tasks and achieves significant speedup compared to previous heuristic approches of adapting auxiliary task weights.

****
Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders
Emile Mathieu, Charline Le Lan, Chris J. Maddison, Ryota Tomioka, Yee Whye Teh
The Variational Auto-Encoder (VAE) is a popular method for learning a generative model and embeddings of the data. Many real datasets are hierarchically structured. However, traditional VAEs map data in a Euclidean latent space which cannot efficiently embed tree-like structures. Hyperbolic spaces with negative curvature can. We therefore endow VAEs with a Poincaré ball model of hyperbolic geometry as a latent space and rigorously derive the necessary methods to work with two main Gaussian generalisations on that space. We empirically show better generalisation to unseen data than the Euclidean counterpart, and can qualitatively and quantitatively better recover hierarchical structures.
****
Training Image Estimators without Image Ground Truth
Zhihao Xia, Ayan Chakrabarti
Deep neural networks have been very successful in compressive-sensing and image restoration applications, as a means to estimate images from partial, blurry, or otherwise degraded measurements. These networks are trained on a large number of corresponding pairs of measurements and ground-truth images, and thus implicitly learn to exploit domain-specific image statistics. But unlike measurement data, it is often expensive or impractical to collect a large training set of ground-truth images in many application settings. In this paper, we introduce an unsupervised framework for training image estimation networks, from a training set that contains only measurements---with two varied measurements per image---but no ground-truth for the full images desired as output. We demonstrate that our framework can be applied for both regular and blind image estimation tasks, where in the latter case parameters of the measurement model (e.g., the blur kernel) are unknown: during inference, and potentially, also during training. We evaluate our framework for training networks for compressive-sensing and blind deconvolution, considering both non-blind and blind training for the latter. Our framework yields models that are nearly as accurate as those from fully supervised training, despite not having access to any ground-truth images.
****
On the Convergence Rate of Training Recurrent Neural Networks
Zeyuan Allen-Zhu, Yuanzhi Li, Zhao Song
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.
****
Minimizers of the Empirical Risk and Risk Monotonicity
Marco Loog, Tom Viering, Alexander Mey
Plotting a learner's average performance against the number of training samples results in a learning curve. Studying such curves on one or more data sets is a way to get to a better understanding of the generalization properties of this learner. The behavior of learning curves is, however, not very well understood and can display (for most researchers) quite unexpected behavior. Our work introduces the formal notion of risk monotonicity, which asks the risk to not deteriorate with increasing training set sizes in expectation over the training samples. We then present the surprising result that various standard learners, specifically those that minimize the empirical risk, can act nonmonotonically irrespective of the training sample size. We provide a theoretical underpinning for specific instantiations from classification, regression, and density estimation. Altogether, the proposed monotonicity notion opens up a whole new direction of research.
****
Factor Group-Sparse Regularization for Efficient Low-Rank Matrix Recovery
Jicong Fan, Lijun Ding, Yudong Chen, Madeleine Udell
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.
****
Möbius Transformation for Fast Inner Product Search on Graph
Zhixin Zhou, Shulong Tan, Zhaozhuo Xu, Ping Li
We present a fast search on graph algorithm for Maximum Inner Product Search (MIPS). This optimization problem is challenging since traditional Approximate Nearest Neighbor (ANN) search methods may not perform efficiently in the non-metric similarity measure. Our proposed method is based on the property that Möbius transformation introduces an isomorphism between a subgraph of l^2-Delaunay graph and Delaunay graph for inner product. Under this observation, we propose a simple but novel graph indexing and searching algorithm to find the optimal solution with the largest inner product with the query. Experiments show our approach leads to significant improvements compared to existing methods.
****
The Label Complexity of Active Learning from Observational Data
Songbai Yan, Kamalika Chaudhuri, Tara Javidi
Counterfactual learning from observational data involves learning a classifier on an entire population based on data that is observed conditioned on a selection policy. This work considers this problem in an active setting, where the learner additionally has access to unlabeled examples and can choose to get a subset of these labeled by an oracle.
****
Hyperbolic Graph Neural Networks
Qi Liu, Maximilian Nickel, Douwe Kiela
Learning from graph-structured data is an important task in machine learning and artificial intelligence, for which Graph Neural Networks (GNNs) have shown great promise. Motivated by recent advances in geometric representation learning, we propose a novel GNN architecture for learning representations on Riemannian manifolds with differentiable exponential and logarithmic maps. We develop a scalable algorithm for modeling the structural properties of graphs, comparing Euclidean and hyperbolic geometry. In our experiments, we show that hyperbolic GNNs can lead to substantial improvements on various benchmark datasets.
****
Learning Fairness in Multi-Agent Systems
Jiechuan Jiang, Zongqing Lu
Fairness is essential for human society, contributing to stability and productivity. Similarly, fairness is also the key for many multi-agent systems. Taking fairness into multi-agent learning could help multi-agent systems become both efficient and stable. However, learning efficiency and fairness simultaneously is a complex, multi-objective, joint-policy optimization. To tackle these difficulties, we propose FEN, a novel hierarchical reinforcement learning model. We first decompose fairness for each agent and propose fair-efficient reward that each agent learns its own policy to optimize. To avoid multi-objective conflict, we design a hierarchy consisting of a controller and several sub-policies, where the controller maximizes the fair-efficient reward by switching among the sub-policies that provides diverse behaviors to interact with the environment. FEN can be trained in a fully decentralized way, making it easy to be deployed in real-world applications. Empirically, we show that FEN easily learns both fairness and efficiency and significantly outperforms baselines in a variety of multi-agent scenarios.
****
On Robustness to Adversarial Examples and Polynomial Optimization
Pranjal Awasthi, Abhratanu Dutta, Aravindan Vijayaraghavan
We study the design of computationally efficient algorithms with provable guarantees, that are robust to adversarial (test time) perturbations. While there has been an explosion of recent work on this topic due to its connections to test time robustness of deep networks, there is limited theoretical understanding of several basic questions like (i) when and how can one design provably robust learning algorithms? (ii) what is the price of achieving robustness to adversarial ex

amples in a computationally efficient manner?
****
In-Place Zero-Space Memory Protection for CNN
Hui Guan, Lin Ning, Zhen Lin, Xipeng Shen, Huiyang Zhou, Seung-Hwan Lim
Convolutional Neural Networks (CNN) are being actively explored for safety-critical applications such as autonomous vehicles and aerospace, where it is essential to ensure the reliability of inference results in the presence of possible memory faults. Traditional methods such as error correction codes (ECC) and Triple Modular Redundancy (TMR) are CNN-oblivious and incur substantial memory overhead and energy cost. This paper introduces in-place zero-space ECC assisted with a new training scheme weight distribution-oriented training. The new method provides the first known zero space cost memory protection for CNNs without compromising the reliability offered by traditional ECC.
****
Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs
Max Simchowitz, Kevin G. Jamieson
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.
****
Discovery of Useful Questions as Auxiliary Tasks
Vivek Veeriah, Matteo Hessel, Zhongwen Xu, Janarthanan Rajendran, Richard L. Lewis, Junhyuk Oh, Hado P. van Hasselt, David Silver, Satinder Singh
Arguably, intelligent agents ought to be able to discover their own questions so that in learning answers for them they learn unanticipated useful knowledge and skills; this departs from the focus in much of machine learning on agents learning answers to externally defined questions. We present a novel method for a reinforcement learning (RL) agent to discover questions formulated as general value functions or GVFs, a fairly rich form of knowledge representation. Specifically, our method uses non-myopic meta-gradients to learn GVF-questions such that learning answers to them, as an auxiliary task, induces useful representations for the main task faced by the RL agent. We demonstrate that auxiliary tasks based on the discovered GVFs are sufficient, on their own, to build representations that support main task learning, and that they do so better than popular hand-designed auxiliary tasks from the literature. Furthermore, we show, in the context of Atari2600 videogames, how such auxiliary tasks, meta-learned alongside the main task, can improve the data efficiency of an actor-critic agent.
****
Sequential Neural Processes
Gautam Singh, Jaesik Yoon, Youngsung Son, Sungjin Ahn
Neural Processes combine the strengths of neural networks and Gaussian processes to achieve both flexible learning and fast prediction in stochastic processes. However, a large class of problems comprise underlying temporal dependency structures in a sequence of stochastic processes that Neural Processes (NP) do not explicitly consider. In this paper, we propose Sequential Neural Processes (SNP) which incorporates a temporal state-transition model of stochastic processes and thus extends its modeling capabilities to dynamic stochastic processes. In applying SNP to dynamic 3D scene modeling, we introduce the Temporal Generative Query Networks. To our knowledge, this is the first 4D model that can deal with the temporal dynamics of 3D scenes. In experiments, we evaluate the proposed methods in dynamic (non-stationary) regression and 4D scene inference and rendering.
****
Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask
Hattie Zhou, Janice Lan, Rosanne Liu, Jason Yosinski
The recent "Lottery Ticket Hypothesis" paper by Frankle & Carbin showed that a simple approach to creating sparse networks (keep the large weights) results in models that are trainable from scratch, but only when starting from the same initial weights. The performance of these networks often exceeds the performance of the non-sparse base model, but for reasons that were not well understood. In thi

s paper we study the three critical components of the Lottery Ticket (LT) algori
thm, showing that each may be varied significantly without impacting the overall
 results. Ablating these factors leads to new insights for why LT networks perfo
rm as well as they do. We show why setting weights to zero is important, how sig
ns are all you need to make the re-initialized network train, and why masking be
haves like training. Finally, we discover the existence of Supermasks, or masks
that can be applied to an untrained, randomly initialized network to produce a m
odel with performance far better than chance (86% on MNIST, 41% on CIFAR-10).
****
Fast and Flexible Multi-Task Classification using Conditional Neural Adaptive Pr
ocesses
James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, Richard E. T
urner
The goal of this paper is to design image classification systems that, after an
initial multi-task training phase, can automatically adapt to new tasks encounte
red at test time. We introduce a conditional neural process based approach to th
e multi-task classification setting for this purpose, and establish connections
to the meta- and few-shot learning literature. The resulting approach, called CN
APs, comprises a classifier whose parameters are modulated by an adaptation netw
ork that takes the current task's dataset as input. We demonstrate that CNAPs ac
hieves state-of-the-art results on the challenging Meta-Dataset benchmark indica
ting high-quality transfer-learning. We show that the approach is robust, avoidi
ng both over-fitting in low-shot regimes and under-fitting in high-shot regimes.
 Timing experiments reveal that CNAPs is computationally efficient at test-time
as it does not involve gradient based adaptation. Finally, we show that trained
models are immediately deployable to continual learning and active learning wher
e they can outperform existing approaches that do not leverage transfer learning
.
****
A Simple Baseline for Bayesian Uncertainty in Deep Learning
Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, Andrew Gordon
 Wilson
We propose SWA-Gaussian (SWAG), a simple, scalable, and general purpose approa
ch for uncertainty representation and calibration in deep learning. Stochastic W
eight Averaging (SWA), which computes the first moment of stochastic gradient de
scent (SGD) iterates with a modified learning rate schedule, has recently been s
hown to improve generalization in deep learning. With SWAG, we fit a Gaussian us
ing the SWA solution as the first moment and a low rank plus diagonal covariance
 also derived from the SGD iterates, forming an approximate posterior distributi
on over neural network weights; we then sample from this Gaussian distribution t
o perform Bayesian model averaging. We empirically find that SWAG approximates t
he shape of the true posterior, in accordance with results describing the statio
nary distribution of SGD iterates. Moreover, we demonstrate that SWAG performs w
ell on a wide variety of tasks, including out of sample detection, calibration,
 and transfer learning, in comparison to many popular alternatives including va
riational inference, MC dropout, KFAC Laplace, and temperature scaling.
****
CPM-Nets: Cross Partial Multi-View Networks
Changqing Zhang, Zongbo Han, yajie cui, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu
Despite multi-view learning progressed fast in past decades, it is still challen
ging due to the difficulty in modeling complex correlation among different views
, especially under the context of view missing. To address the challenge, we pro
pose a novel framework termed Cross Partial Multi-View Networks (CPM-Nets). In t
his framework, we first give a formal definition of completeness and versatility
 for multi-view representation and then theoretically prove the versatility of t
he latent representation learned from our algorithm. To achieve the completeness
, the task of learning latent multi-view representation is specifically translat
ed to degradation process through mimicking data transmitting, such that the opt
imal tradeoff between consistence and complementarity across different views cou
ld be achieved. In contrast with methods that either complete missing views or g

roup samples according to view-missing patterns, our model fully exploits all samples and all views to produce structured representation for interpretability. Extensive experimental results validate the effectiveness of our algorithm over existing state-of-the-arts.
****
Low-Complexity Nonparametric Bayesian Online Prediction with Universal Guarantees
Alix LHERITIER, Frederic Cazals
We propose a novel nonparametric online predictor for discrete labels conditioned on multivariate continuous features. The predictor is based on a feature space discretization induced by a full-fledged k-d tree with randomly picked directions and a recursive Bayesian distribution, which allows to automatically learn the most relevant feature scales characterizing the conditional distribution. We prove its pointwise universality, i.e., it achieves a normalized log loss performance asymptotically as good as the true conditional entropy of the labels given the features. The time complexity to process the n-th sample point is O(log n) in probability with respect to the distribution generating the data points, whereas other exact nonparametric methods require to process all past observations. Experiments on challenging datasets show the computational and statistical efficiency of our algorithm in comparison to standard and state-of-the-art methods.
****
Finding the Needle in the Haystack with Convolutions: on the benefits of architectural bias
Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, Joan Bruna
Despite the phenomenal success of deep neural networks in a broad range of learning tasks, there is a lack of theory to understand the way they work. In particular, Convolutional Neural Networks (CNNs) are known to perform much better than Fully-Connected Networks (FCNs) on spatially structured data: the architectural structure of CNNs benefits from prior knowledge on the features of the data, for instance their translation invariance. The aim of this work is to
understand this fact through the lens of dynamics in the loss landscape.
****
Efficiently avoiding saddle points with zero order methods: No gradients required
Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Georgios Piliouras
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 metrics for persistence-based summaries and applications for graph classification
Qi Zhao, Yusu Wang
Recently a new feature representation and data analysis methodology based on a topological tool called persistent homology (and its persistence diagram summary) has gained much momentum. A series of methods have been developed to map a persistence diagram to a vector representation so as to facilitate the downstream use of machine learning tools. In these approaches, the importance (weight) of different persistence features are usually pre-set. However often in practice, the choice of the weight-function should depend on the nature of the specific data at hand. It is thus highly desirable to learn a best weight-function (and thus metric for persistence diagrams) from labelled data. We study this problem and develop a new weighted kernel, called WKPI, for persistence summaries, as well as an optimization framework to learn the weight (and thus kernel). We apply the learned kernel to the challenging task of graph classification, and show that our WKPI-based classification framework obtains similar or (sometimes significantly) better results than the best results from a range of previous graph classification frameworks on a collection of benchmark datasets.
****
PasteGAN: A Semi-Parametric Method to Generate Image from Scene Graph

Yikang LI, Tao Ma, Yeqi Bai, Nan Duan, Sining Wei, Xiaogang Wang
Despite some exciting progress on high-quality image generation from structured (scene graphs) or free-form (sentences) descriptions, most of them only guarantee the image-level semantical consistency, i.e. the generated image matching the semantic meaning of the description. They still lack the investigations on synthesizing the images in a more controllable way, like finely manipulating the visual appearance of every object. Therefore, to generate the images with preferred objects and rich interactions, we propose a semi-parametric method, PasteGAN, for generating the image from the scene graph and the image crops, where spatial arrangements of the objects and their pair-wise relationships are defined by the scene graph and the object appearances are determined by the given object crops. To enhance the interactions of the objects in the output, we design a Crop Refining Network and an Object-Image Fuser to embed the objects as well as their relationships into one map. Multiple losses work collaboratively to guarantee the generated images highly respecting the crops and complying with the scene graphs while maintaining excellent image quality. A crop selector is also proposed to pick the most-compatible crops from our external object tank by encoding the interactions around the objects in the scene graph if the crops are not provided. Evaluated on Visual Genome and COCO-Stuff dataset, our proposed method significantly outperforms the SOTA methods on Inception Score, Diversity Score and Fre■chet Inception Distance. Extensive experiments also demonstrate our method’s ability to generate complex and diverse images with given objects. The code is available at https://github.com/yikang-li/PasteGAN.
****
Learning Local Search Heuristics for Boolean Satisfiability
Emre Yolcu, Barnabas Poczos
We present an approach to learn SAT solver heuristics from scratch through deep reinforcement learning with a curriculum. In particular, we incorporate a graph neural network in a stochastic local search algorithm to act as the variable selection heuristic. We consider Boolean satisfiability problems from different classes and learn specialized heuristics for each class. Although we do not aim to compete with the state-of-the-art SAT solvers in run time, we demonstrate that the learned heuristics allow us to find satisfying assignments in fewer steps compared to a generic heuristic, and we provide analysis of our results through experiments.
****
Learning to Perform Local Rewriting for Combinatorial Optimization
Xinyun Chen, Yuandong Tian
Search-based methods for hard combinatorial optimization are often guided by heuristics. Tuning heuristics in various conditions and situations is often time-consuming. In this paper, we propose NeuRewriter that learns a policy to pick heuristics and rewrite the local components of the current solution to iteratively improve it until convergence. The policy factorizes into a region-picking and a rule-picking component, each parameterized by a neural network trained with actor-critic methods in reinforcement learning. NeuRewriter captures the general structure of combinatorial problems and shows strong performance in three versatile tasks: expression simplification, online job scheduling and vehicle routing problems. NeuRewriter outperforms the expression simplification component in Z3; outperforms DeepRM and Google OR-tools in online job scheduling; and outperforms recent neural baselines and Google OR-tools in vehicle routing problems.
****
A Unified Bellman Optimality Principle Combining Reward Maximization and Empowerment
Felix Leibfried, Sergio Pascual-Díaz, Jordi Grau-Moya
Empowerment is an information-theoretic method that can be used to intrinsically motivate learning agents. It attempts to maximize an agent's control over the environment by encouraging visiting states with a large number of reachable next states. Empowered learning has been shown to lead to complex behaviors, without requiring an explicit reward signal. In this paper, we investigate the use of empowerment in the presence of an extrinsic reward signal. We hypothesize that emp

owerment can guide reinforcement learning (RL) agents to find good early behavioral solutions by encouraging highly empowered states.

We propose a unified Bellman optimality principle for empowered reward maximization. Our empowered reward maximization approach generalizes both Bellman's optimality principle as well as recent information-theoretical extensions to it. We prove uniqueness of the empowered values and show convergence to the optimal solution. We then apply this idea to develop off-policy actor-critic RL algorithms which we validate in high-dimensional continuous robotics domains (MuJoCo). Our methods demonstrate improved initial and competitive final performance compared to model-free state-of-the-art techniques.

****

Learning Representations for Time Series Clustering

Qianli Ma, Jiawei Zheng, Sen Li, Gary W. Cottrell

Time series clustering is an essential unsupervised technique in cases when category information is not available. It has been widely applied to genome data, anomaly detection, and in general, in any domain where pattern detection is important. Although feature-based time series clustering methods are robust to noise and outliers, and can reduce the dimensionality of the data, they typically rely on domain knowledge to manually construct high-quality features. Sequence to sequence (seq2seq) models can learn representations from sequence data in an unsupervised manner by designing appropriate learning objectives, such as reconstruction and context prediction. When applying seq2seq to time series clustering, obtaining a representation that effectively represents the temporal dynamics of the sequence, multi-scale features, and good clustering properties remains a challenge. How to best improve the ability of the encoder is still an open question. Here we propose a novel unsupervised temporal representation learning model, named Deep Temporal Clustering Representation (DTCR), which integrates the temporal reconstruction and K-means objective into the seq2seq model. This approach leads to improved cluster structures and thus obtains cluster-specific temporal representations. Also, to enhance the ability of encoder, we propose a fake-sample generation strategy and auxiliary classification task. Experiments conducted on extensive time series datasets show that DTCR is state-of-the-art compared to existing methods. The visualization analysis not only shows the effectiveness of cluster-specific representation but also shows the learning process is robust, even if K-means makes mistakes.

****

Statistical-Computational Tradeoff in Single Index Models

Lingxiao Wang, Zhuoran Yang, Zhaoran 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.

****

Probabilistic Logic Neural Networks for Reasoning

Meng Qu, Jian Tang

Knowledge graph reasoning, which aims at predicting missing facts through reasoning with observed facts, is critical for many applications. Such a problem has been widely explored by traditional logic rule-based approaches and recent knowledge graph embedding methods. A principled logic rule-based approach is the Markov Logic Network (MLN), which is able to leverage domain knowledge with first-order logic and meanwhile handle uncertainty. However, the inference in MLNs is usually very difficult due to the complicated graph structures. Different from MLNs, knowledge graph embedding methods (e.g. TransE, DistMult) learn effective entity and relation embeddings for reasoning, which are much more effective and efficient. However, they are unable to leverage domain knowledge. In this paper, we propose the probabilistic Logic Neural Network (pLogicNet), which combines the advantages of both methods. A pLogicNet defines the joint distribution of all possible triplets by using a Markov logic network with first-order logic, which can be efficiently optimized with the variational EM algorithm. Specifically, in the E-step, a knowledge graph embedding model is used for inferring the missing tr

iplets, while in the M-step, the weights of the logic rules are updated accordin g to both the observed and predicted triplets. Experiments on multiple knowledge graphs prove the effectiveness of pLogicNet over many competitive baselines.
****
Joint-task Self-supervised Learning for Temporal Correspondence
Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, Ming-Hsuan Ya ng
This paper proposes to learn reliable dense correspondence from videos in a self -supervised manner. Our learning process integrates two highly related tasks: tr acking large image regions and establishing fine-grained pixel-level association s between consecutive video frames. We exploit the synergy between both tasks th rough a shared inter-frame affinity matrix, which simultaneously models transiti ons between video frames at both the region- and pixel-levels. While region-leve l localization helps reduce ambiguities in fine-grained matching by narrowing do wn search regions; fine-grained matching provides bottom-up features to facilita te region-level localization. Our method outperforms the state-of-the-art self-s upervised methods on a variety of visual correspondence tasks, including video-o bject and part-segmentation propagation, keypoint tracking, and object tracking. Our self-supervised method even surpasses the fully-supervised affinity feature representation obtained from a ResNet-18 pre-trained on the ImageNet.
****
Learning Sparse Distributions using Iterative Hard Thresholding
Jacky Y. Zhang, Rajiv Khanna, Anastasios Kyrillidis, Oluwasanmi O. Koyejo
Iterative hard thresholding (IHT) is a projected gradient descent algorithm, kno wn to achieve state of the art performance for a wide range of structured estima tion problems, such as sparse inference. In this work, we consider IHT as a solu tion to the problem of learning sparse discrete distributions. We study the hard ness of using IHT on the space of measures. As a practical alternative, we propo se a greedy approximate projection which simultaneously captures appropriate not ions of sparsity in distributions, while satisfying the simplex constraint, and investigate the convergence behavior of the resulting procedure in various setti ngs. Our results show, both in theory and practice, that IHT can achieve state o f the art results for learning sparse distributions.
****
On Distributed Averaging for Stochastic k-PCA
Aditya Bhaskara, Pruthuvi Maheshakya Wijewardena
In the stochastic k-PCA problem, we are given i.i.d. samples from an unknown dis tribution over vectors, and the goal is to compute the top k eigenvalues and eig envectors of the moment matrix. In the simplest distributed variant, we have 'm' machines each of which receives 'n' samples. Each machine performs some computa tion and sends an O(k) size summary of the local dataset to a central server. Th e server performs an aggregation and computes the desired eigenvalues and vector s. The goal is to achieve the same effect as the server computing using m*n samp les by itself. The main choices in this framework are the choice of the summary, and the method of aggregation. We consider a slight variant of the well-studied "distributed averaging" approach, and prove that this leads to significantly be tter bounds on the dependence between 'n' and the eigenvalue gaps. Our method ca n also be applied directly to a setting where the "right" value of the parameter k (i.e., one for which there is a non-trivial eigenvalue gap) is not known exac tly. This is a common issue in practice which prior methods were unable to addre ss.
****
Learning dynamic polynomial proofs
Alhussein Fawzi, Mateusz Malinowski, Hamza Fawzi, Omar Fawzi
Polynomial inequalities lie at the heart of many mathematical disciplines. In th is paper, we consider the fundamental computational task of automatically search ing for proofs of polynomial inequalities. We adopt the framework of semi-algebr aic proof systems that manipulate polynomial inequalities via elementary inferen ce rules that infer new inequalities from the premises. These proof systems are known to be very powerful, but searching for proofs remains a major difficulty.

In this work, we introduce a machine learning based method to search for a dynam
ic proof within these proof systems. We propose a deep reinforcement learning fr
amework that learns an embedding of the polynomials and guides the choice of inf
erence rules, taking the inherent symmetries of the problem as an inductive bias
. We compare our approach with powerful and widely-studied linear programming hi
erarchies based on static proof systems, and show that our method reduces the s
ize of the linear program by several orders of magnitude while also improving pe
rformance. These results hence pave the way towards augmenting powerful and well
-studied semi-algebraic proof systems with machine learning guiding strategies f
or enhancing the expressivity of such proof systems.
****
Efficient Communication in Multi-Agent Reinforcement Learning via Variance Based
 Control
Sai Qian Zhang, Qi Zhang, Jieyu Lin
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.
****
Global Convergence of Gradient Descent for Deep Linear Residual Networks
Lei Wu, Qingcan Wang, Chao Ma
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.
****
Dying Experts: Efficient Algorithms with Optimal Regret Bounds
Hamid Shayestehmanesh, Sajjad Azami, Nishant A. Mehta
We study a variant of decision-theoretic online learning in which the set of exp
erts that are available to Learner can shrink over time. This is a restricted ve
rsion of the well-studied sleeping experts problem, itself a generalization of t
he fundamental game of prediction with expert advice. Similar to many works in t
his direction, our benchmark is the ranking regret. Various results suggest that
 achieving optimal regret in the fully adversarial sleeping experts problem is c
omputationally hard. This motivates our relaxation where any expert that goes to
 sleep will never again wake up. We call this setting "dying experts" and study
it in two different cases: the case where the learner knows the order in which t
he experts will die and the case where the learner does not. In both cases, we p
rovide matching upper and lower bounds on the ranking regret in the fully advers
arial setting. Furthermore, we present new, computationally efficient algorithms
 that obtain our optimal upper bounds.
****
A Bayesian Theory of Conformity in Collective Decision Making
Koosha Khalvati, Saghar Mirbagheri, Seongmin A. Park, Jean-Claude Dreher, Rajesh
 PN Rao
In collective decision making, members of a group need to coordinate their actio
ns in order to achieve a desirable outcome. When there is no direct communicatio
n between group members, one should decide based on inferring others' intentions
 from their actions. The inference of others' intentions is called "theory of mi
nd" and can involve different levels of reasoning, from a single inference on a
hidden variable to considering others partially or fully optimal and reasoning a
bout their actions conditioned on one's own actions (levels of “theory of mind”)
. In this paper, we present a new Bayesian theory of collective decision making
based on a simple yet most commonly observed behavior: conformity. We show that
such a Bayesian framework allows one to achieve any level of theory of mind in c
ollective decision making. The viability of our framework is demonstrated on two
 different experiments, a consensus task with 120 subjects and a volunteer's dil
emma task with 29 subjects, each with multiple conditions.
****
Poisson-Randomized Gamma Dynamical Systems

Aaron Schein, Scott Linderman, Mingyuan Zhou, David Blei, Hanna Wallach
This paper presents the Poisson-randomized gamma dynamical system (PRGDS), a model for sequentially observed count tensors that encodes a strong inductive bias toward sparsity and burstiness. The PRGDS is based on a new motif in Bayesian latent variable modeling, an alternating chain of discrete Poisson and continuous gamma latent states that is analytically convenient and computationally tractable. This motif yields closed-form complete conditionals for all variables by way of the Bessel distribution and a novel discrete distribution that we call the shifted confluent hypergeometric distribution. We draw connections to closely related models and compare the PRGDS to these models in studies of real-world count data sets of text, international events, and neural spike trains. We find that a sparse variant of the PRGDS, which allows the continuous gamma latent states to take values of exactly zero, often obtains better predictive performance than other models and is uniquely capable of inferring latent structures that are highly localized in time.
****
Bayesian Layers: A Module for Neural Network Uncertainty
Dustin Tran, Mike Dusenberry, Mark van der Wilk, Danijar Hafner
We describe Bayesian Layers, a module designed for fast experimentation with neural network uncertainty. It extends neural network libraries with drop-in replacements for common layers. This enables composition via a unified abstraction over deterministic and stochastic functions and allows for scalability via the underlying system. These layers capture uncertainty over weights (Bayesian neural nets), pre-activation units (dropout), activations (stochastic output layers''), or the function itself (Gaussian processes). They can also be reversible to propagate uncertainty from input to output. We include code examples for common architectures such as Bayesian LSTMs, deep GPs, and flow-based models. As demonstration, we fit a 5-billion parameterBayesian Transformer'' on 512 TPUv2 cores for uncertainty in machine translation and a Bayesian dynamics model for model-based planning. Finally, we show how Bayesian Layers can be used within the Edward2 language for probabilistic programming with stochastic processes.
****
Sequence Modeling with Unconstrained Generation Order
Dmitrii Emelianenko, Elena Voita, Pavel Serdyukov
The dominant approach to sequence generation is to produce a sequence in some predefined order, e.g. left to right. In contrast, we propose a more general model that can generate the output sequence by inserting tokens in any arbitrary order. Our model learns decoding order as a result of its training procedure. Our experiments show that this model is superior to fixed order models on a number of sequence generation tasks, such as Machine Translation, Image-to-LaTeX and Image Captioning.
****
Online Continual Learning with Maximal Interfered Retrieval
Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, Lucas Page-Caccia
Continual learning, the setting where a learning agent is faced with a never-ending stream of data, continues to be a great challenge for modern machine learning systems. In particular the online or "single-pass through the data" setting has gained attention recently as a natural setting that is difficult to tackle. Methods based on replay, either generative or from a stored memory, have been shown to be effective approaches for continual learning, matching or exceeding the state of the art in a number of standard benchmarks. These approaches typically rely on randomly selecting samples from the replay memory or from a generative model, which is suboptimal. In this work, we consider a controlled sampling of memories for replay. We retrieve the samples which are most interfered, i.e. whose prediction will be most negatively impacted by the foreseen parameters update. We show a formulation for this sampling criterion in both the generative replay and the experience replay setting, producing consistent gains in performance and greatly reduced forgetting. We release an implementation of our method at https://github.com/optimass/MaximallyInterferedRetrieval

****
Visualizing and Measuring the Geometry of BERT
Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viegas, Andy Coenen, Adam Pearce, Been Kim
Transformer architectures show significant promise for natural language processing. Given that a single pretrained model can be fine-tuned to perform well on many different tasks, these networks appear to extract generally useful linguistic features. A natural question is how such networks represent this information internally. This paper describes qualitative and quantitative investigations of one particularly effective model, BERT. At a high level, linguistic features seem to be represented in separate semantic and syntactic subspaces. We find evidence of a fine-grained geometric representation of word senses. We also present empirical descriptions of syntactic representations in both attention matrices and individual word embeddings, as well as a mathematical argument to explain the geometry of these representations.
****
Learning to Predict Without Looking Ahead: World Models Without Forward Prediction
Daniel Freeman, David Ha, Luke Metz
Much of model-based reinforcement learning involves learning a model of an agent's world, and training an agent to leverage this model to perform a task more efficiently. While these models are demonstrably useful for agents, every naturally occurring model of the world of which we are aware---e.g., a brain---arose as the byproduct of competing evolutionary pressures for survival, not minimization of a supervised forward-predictive loss via gradient descent. That useful models can arise out of the messy and slow optimization process of evolution suggests that forward-predictive modeling can arise as a side-effect of optimization under the right circumstances. Crucially, this optimization process need not explicitly be a forward-predictive loss. In this work, we introduce a modification to traditional reinforcement learning which we call observational dropout, whereby we limit the agents ability to observe the real environment at each timestep. In doing so, we can coerce an agent into learning a world model to fill in the observation gaps during reinforcement learning. We show that the emerged world model, while not explicitly trained to predict the future, can help the agent learn key skills required to perform well in its environment. Videos of our results available at https://learningtopredict.github.io/
****
Deep Generalized Method of Moments for Instrumental Variable Analysis
Andrew Bennett, Nathan Kallus, Tobias Schnabel
Instrumental variable analysis is a powerful tool for estimating causal effects when randomization or full control of confounders is not possible. The application of standard methods such as 2SLS, GMM, and more recent variants are significantly impeded when the causal effects are complex, the instruments are high-dimensional, and/or the treatment is high-dimensional. In this paper, we propose the DeepGMM algorithm to overcome this. Our algorithm is based on a new variational reformulation of GMM with optimal inverse-covariance weighting that allows us to efficiently control very many moment conditions. We further develop practical techniques for optimization and model selection that make it particularly successful in practice. Our algorithm is also computationally tractable and can handle large-scale datasets. Numerical results show our algorithm matches the performance of the best tuned methods in standard settings and continues to work in high-dimensional settings where even recent methods break.
****
Copulas as High-Dimensional Generative Models: Vine Copula Autoencoders
Natasa Tagasovska, Damien Ackerer, Thibault Vatter
We introduce the vine copula autoencoder (VCAE), a flexible generative model for high-dimensional distributions built in a straightforward three-step procedure. First, an autoencoder (AE) compresses the data into a lower dimensional representation.
Second, the multivariate distribution of the encoded data is estimated with vine

copulas.
Third, a generative model is obtained by combining the estimated distribution with the decoder part of the AE.
As such, the proposed approach can transform any already trained AE into a flexible generative model at a low computational cost.
This is an advantage over existing generative models such as adversarial networks and variational AEs which can be difficult to train and can impose strong assumptions on the latent space.
Experiments on MNIST, Street View House Numbers and Large-Scale CelebFaces Attributes datasets show that VCAEs can achieve competitive results to standard baselines.
****
Implicit Semantic Data Augmentation for Deep Networks
Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, Cheng Wu
In this paper, we propose a novel implicit semantic data augmentation (ISDA) approach to complement traditional augmentation techniques like flipping, translation or rotation. Our work is motivated by the intriguing property that deep networks are surprisingly good at linearizing features, such that certain directions in the deep feature space correspond to meaningful semantic transformations, e.g., adding sunglasses or changing backgrounds. As a consequence, translating training samples along many semantic directions in the feature space can effectively augment the dataset to improve generalization. To implement this idea effectively and efficiently, we first perform an online estimate of the covariance matrix of deep features for each class, which captures the intra-class semantic variations. Then random vectors are drawn from a zero-mean normal distribution with the estimated covariance to augment the training data in that class. Importantly, instead of augmenting the samples explicitly, we can directly minimize an upper bound of the expected cross-entropy (CE) loss on the augmented training set, leading to a highly efficient algorithm. In fact, we show that the proposed ISDA amounts to minimizing a novel robust CE loss, which adds negligible extra computational cost to a normal training procedure. Although being simple, ISDA consistently improves the generalization performance of popular deep models (ResNets and DenseNets) on a variety of datasets, e.g., CIFAR-10, CIFAR-100 and ImageNet. Code for reproducing our results are available at https://github.com/blackfeather-wang/ISDA-for-Deep-Networks.
****
q-means: A quantum algorithm for unsupervised machine learning
Iordanis Kerenidis, Jonas Landman, Alessandro Luongo, Anupam Prakash
Quantum information is a promising new paradigm for fast computations that can provide substantial speedups for many algorithms we use today. Among them, quantum machine learning is one of the most exciting applications of quantum computers. In this paper, we introduce q-means, a new quantum algorithm for clustering. It is a quantum version of a robust k-means algorithm, with similar convergence and precision guarantees. We also design a method to pick the initial centroids equivalent to the classical k-means++ method. Our algorithm provides currently an exponential speedup in the number of points of the dataset, compared to the classical k-means algorithm. We also detail the running time of q-means when applied to well-clusterable datasets. We provide a detailed runtime analysis and numerical simulations for specific datasets. Along with the algorithm, the theorems and tools introduced in this paper can be reused for various applications in quantum machine learning.
****
RUDDER: Return Decomposition for Delayed Rewards
Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, Sepp Hochreiter
We propose RUDDER, a novel reinforcement learning approach for delayed rewards in finite Markov decision processes (MDPs). In MDPs the Q-values are equal to the expected immediate reward plus the expected future rewards. The latter are related to bias problems in temporal difference (TD) learning and to high variance problems in Monte Carlo (MC) learning. Both problems are even more severe when re

wards are delayed. RUDDER aims at making the expected future rewards zero, which simplifies Q-value estimation to computing the mean of the immediate reward. We propose the following two new concepts to push the expected future rewards towa rd zero. (i) Reward redistribution that leads to return-equivalent decision proc esses with the same optimal policies and, when optimal, zero expected future rew ards. (ii) Return decomposition via contribution analysis which transforms the r einforcement learning task into a regression task at which deep learning excels. On artificial tasks with delayed rewards, RUDDER is significantly faster than M C and exponentially faster than Monte Carlo Tree Search (MCTS), TD(λ), and rewar d shaping approaches. At Atari games, RUDDER on top of a Proximal Policy Optimiz ation (PPO) baseline improves the scores, which is most prominent at games with delayed rewards.
****
Learning-Based Low-Rank Approximations
Piotr Indyk, Ali Vakilian, Yang Yuan
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.
****
Convergence Guarantees for Adaptive Bayesian Quadrature Methods
Motonobu Kanagawa, Philipp Hennig
Adaptive Bayesian quadrature (ABQ) is a powerful approach to numerical integrati on that empirically compares favorably with Monte Carlo integration on problems of medium dimensionality (where non-adaptive quadrature is not competitive).
Its key ingredient is an acquisition function that changes as a function of pre viously collected values of the integrand.
While this adaptivity appears to be empirically powerful, it complicates analysi s. Consequently, there are no theoretical guarantees so far for this class of me thods. In this work, for a broad class of adaptive Bayesian quadrature methods, we prove consistency, deriving non-tight but informative convergence rates. To d o so we introduce a new concept we call \emph{weak adaptivity}. Our results iden tify a large and flexible class of adaptive Bayesian quadrature rules as consist ent, within which practitioners can develop empirically efficient methods.
****
A First-Order Algorithmic Framework for Distributionally Robust Logistic Regress ion
JIAJIN LI, SEN HUANG, Anthony Man-Cho So
Wasserstein distance-based distributionally robust optimization (DRO) has receiv ed much attention lately due to its ability to provide a robustness interpretati on of various learning models. Moreover, many of the DRO problems that arise in the learning context admits exact convex reformulations and hence can be tackled by off-the-shelf solvers. Nevertheless, the use of such solvers severely limits the applicability of DRO in large-scale learning problems, as they often rely o n general purpose interior-point algorithms. On the other hand, there are very f ew works that attempt to develop fast iterative methods to solve these DRO probl ems, which typically possess complicated structures. In this paper, we take a fi rst step towards resolving the above difficulty by developing a first-order algo rithmic framework for tackling a class of Wasserstein distance-based distributio nally robust logistic regression (DRLR) problem. Specifically, we propose a nove l linearized proximal ADMM to solve the DRLR problem, whose objective is convex but consists of a smooth term plus two non-separable non-smooth terms. We prove that our method enjoys a sublinear convergence rate. Furthermore, we conduct thr ee different experiments to show its superb performance on both synthetic and re al-world datasets. In particular, our method can achieve the same accuracy up to 800+ times faster than the standard off-the-shelf solver.
****
Theoretical Analysis of Adversarial Learning: A Minimax Approach
Zhuozhuo Tu, Jingwei Zhang, Dacheng Tao
In this paper, we propose a general theoretical method for analyzing the risk bo

und in the presence of adversaries. Specifically, we try to fit the adversarial learning problem into the minimax framework. We first show that the original adversarial learning problem can be transformed into a minimax statistical learning problem by introducing a transport map between distributions. Then, we prove a new risk bound for this minimax problem in terms of covering numbers under a weak version of Lipschitz condition. Our method can be applied to multi-class classification and popular loss functions including the hinge loss and ramp loss. As some illustrative examples, we derive the adversarial risk bounds for SVMs and deep neural networks, and our bounds have two data-dependent terms, which can be optimized for achieving adversarial robustness.
****

Compositional De-Attention Networks
Yi Tay, Anh Tuan Luu, Aston Zhang, Shuohang Wang, Siu Cheung Hui
Attentional models are distinctly characterized by their ability to learn relative importance, i.e., assigning a different weight to input values. This paper proposes a new quasi-attention that is compositional in nature, i.e., learning whether to \textit{add}, \textit{subtract} or \textit{nullify} a certain vector when learning representations. This is strongly contrasted with vanilla attention, which simply re-weights input tokens. Our proposed \textit{Compositional De-Attention} (CoDA) is fundamentally built upon the intuition of both similarity and dissimilarity (negative affinity) when computing affinity scores, benefiting from a greater extent of expressiveness. We evaluate CoDA on six NLP tasks, i.e. open domain question answering, retrieval/ranking, natural language inference, machine translation, sentiment analysis and text2code generation. We obtain promising experimental results, achieving state-of-the-art performance on several tasks/datasets.
****

Robust Attribution Regularization
Jiefeng Chen, Xi Wu, Vaibhav Rastogi, Yingyu Liang, Somesh Jha
An emerging problem in trustworthy machine learning is to train models that produce robust interpretations for their predictions. We take a step towards solving this problem through the lens of axiomatic attribution of neural networks. Our theory is grounded in the recent work, Integrated Gradients (IG) [STY17], in axiomatically attributing a neural network’s output change to its input change. We propose training objectives in classic robust optimization models to achieve robust IG attributions. Our objectives give principled generalizations of previous objectives designed for robust predictions, and they naturally degenerate to classic soft-margin training for one-layer neural networks. We also generalize previous theory and prove that the objectives for different robust optimization models are closely related. Experiments demonstrate the effectiveness of our method, and also point to intriguing problems which hint at the need for better optimization techniques or better neural network architectures for robust attribution training.
****

Semantic-Guided Multi-Attention Localization for Zero-Shot Learning
Yizhe Zhu, Jianwen Xie, Zhiqiang Tang, Xi Peng, Ahmed Elgammal
Zero-shot learning extends the conventional object classification to the unseen class recognition by introducing semantic representations of classes. Existing approaches predominantly focus on learning the proper mapping function for visual-semantic embedding, while neglecting the effect of learning discriminative visual features. In this paper, we study the significance of the discriminative region localization. We propose a semantic-guided multi-attention localization model, which automatically discovers the most discriminative parts of objects for zero-shot learning without any human annotations. Our model jointly learns cooperative global and local features from the whole object as well as the detected parts to categorize objects based on semantic descriptions. Moreover, with the joint supervision of embedding softmax loss and class-center triplet loss, the model is encouraged to learn features with high inter-class dispersion and intra-class compactness. Through comprehensive experiments on three widely used zero-shot learning benchmarks, we show the efficacy of the multi-attention localization and

our proposed approach improves the state-of-the-art results by a considerable margin.
****
Distributionally Robust Optimization and Generalization in Kernel Methods
Matthew Staib, Stefanie Jegelka
Distributionally robust optimization (DRO) has attracted attention in machine learning due to its connections to regularization, generalization, and robustness. Existing work has considered uncertainty sets based on phi-divergences and Wasserstein distances, each of which have drawbacks. In this paper, we study DRO with uncertainty sets measured via maximum mean discrepancy (MMD). We show that MMD DRO is roughly equivalent to regularization by the Hilbert norm and, as a byproduct, reveal deep connections to classic results in statistical learning. In particular, we obtain an alternative proof of a generalization bound for Gaussian kernel ridge regression via a DRO lense. The proof also suggests a new regularizer. Our results apply beyond kernel methods: we derive a generically applicable approximation of MMD DRO, and show that it generalizes recent work on variance-based regularization.
****
Kernel Instrumental Variable Regression
Rahul Singh, Maneesh Sahani, Arthur Gretton
Instrumental variable (IV) regression is a strategy for learning causal relationships in observational data. If measurements of input X and output Y are confounded, the causal relationship can nonetheless be identified if an instrumental variable Z is available that influences X directly, but is conditionally independent of Y given X and the unmeasured confounder. The classic two-stage least squares algorithm (2SLS) simplifies the estimation problem by modeling all relationships as linear functions. We propose kernel instrumental variable regression (KIV), a nonparametric generalization of 2SLS, modeling relations among X, Y, and Z as nonlinear functions in reproducing kernel Hilbert spaces (RKHSs). We prove the consistency of KIV under mild assumptions, and derive conditions under which convergence occurs at the minimax optimal rate for unconfounded, single-stage RKHS regression. In doing so, we obtain an efficient ratio between training sample sizes used in the algorithm's first and second stages. In experiments, KIV outperforms state of the art alternatives for nonparametric IV regression.
****
Metalearned Neural Memory
Tsendsuren Munkhdalai, Alessandro Sordoni, TONG WANG, Adam Trischler
We augment recurrent neural networks with an external memory mechanism that builds upon recent progress in metalearning. We conceptualize this memory as a rapidly adaptable function that we parameterize as a deep neural network. Reading from the neural memory function amounts to pushing an input (the key vector) through the function to produce an output (the value vector). Writing to memory means changing the function; specifically, updating the parameters of the neural network to encode desired information. We leverage training and algorithmic techniques from metalearning to update the neural memory function in one shot. The proposed memory-augmented model achieves strong performance on a variety of learning problems, from supervised question answering to reinforcement learning.
****
Learning Bayesian Networks with Low Rank Conditional Probability Tables
Adarsh Barik, Jean Honorio
In this paper, we provide a method to learn the directed structure of a Bayesian network using data. The data is accessed by making conditional probability queries to a black-box model. We introduce a notion of simplicity of representation of conditional probability tables for the nodes in the Bayesian network, that we call `low rankness''. We connect this notion to the Fourier transformation of real valued set functions and propose a method which learns the exact directed structure of alow rank` Bayesian network using very few queries. We formally prove that our method correctly recovers the true directed structure, runs in polynomial time and only needs polynomial samples with respect to the number of nodes. We also provide further improvements in efficiency if we have access to some obs

ervational data.
****
Large Scale Adversarial Representation Learning
Jeff Donahue, Karen Simonyan
Adversarially trained generative models (GANs) have recently achieved compelling image synthesis results. But despite early successes in using GANs for unsupervised representation learning, they have since been superseded by approaches based on self-supervision. In this work we show that progress in image generation quality translates to substantially improved representation learning performance. Our approach, BigBiGAN, builds upon the state-of-the-art BigGAN model, extending it to representation learning by adding an encoder and modifying the discriminator. We extensively evaluate the representation learning and generation capabilities of these BigBiGAN models, demonstrating that these generation-based models achieve the state of the art in unsupervised representation learning on ImageNet, as well as compelling results in unconditional image generation.
****
Hindsight Credit Assignment
Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P. van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, Remi Munos
We consider the problem of efficient credit assignment in reinforcement learning. In order to efficiently and meaningfully utilize new data, we propose to explicitly assign credit to past decisions based on the likelihood of them having led to the observed outcome. This approach uses new information in hindsight, rather than employing foresight. Somewhat surprisingly, we show that value functions can be rewritten through this lens, yielding a new family of algorithms. We study the properties of these algorithms, and empirically show that they successfully address important credit assignment challenges, through a set of illustrative tasks.
****
Zero-shot Learning via Simultaneous Generating and Learning
Hyeonwoo Yu, Beomhee Lee
To overcome the absence of training data for unseen classes, conventional zero-shot learning approaches mainly train their model on seen datapoints and leverage the semantic descriptions for both seen and unseen classes.
Beyond exploiting relations between classes of seen and unseen, we present a deep generative model to provide the model with experience about both seen and unseen classes.
Based on the variational auto-encoder with class-specific multi-modal prior, the proposed method learns the conditional distribution of seen and unseen classes. In order to circumvent the need for samples of unseen classes, we treat the non-existing data as missing examples.
That is, our network aims to find optimal unseen datapoints and model parameters, by iteratively following the generating and learning strategy.
Since we obtain the conditional generative model for both seen and unseen classes, classification as well as generation can be performed directly without any off-the-shell classifiers.
In experimental results, we demonstrate that the proposed generating and learning strategy makes the model achieve the outperforming results compared to that trained only on the seen classes, and also to the several state-of-the-art methods.
****
Direct Optimization through $\arg \max$ for Discrete Variational Auto-Encoder
Guy Lorberbom, Andreea Gane, Tommi Jaakkola, Tamir Hazan
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.
****
Generalization Error Analysis of Quantized Compressive Learning

Xiaoyun Li, Ping Li
Compressive learning is an effective method to deal with very high dimensional datasets by applying learning algorithms in a randomly projected lower dimensional space. In this paper, we consider the learning problem where the projected data is further compressed by scalar quantization, which is called quantized compressive learning. Generalization error bounds are derived for three models: nearest neighbor (NN) classifier, linear classifier and least squares regression. Besides studying finite sample setting, our asymptotic analysis shows that the inner product estimators have deep connection with NN and linear classification problem through the variance of their debiased counterparts. By analyzing the extra error term brought by quantization, our results provide useful implications to the choice of quantizers in applications involving different learning tasks. Empirical study is also conducted to validate our theoretical findings.
****
Successor Uncertainties: Exploration and Uncertainty in Temporal Difference Learning
David Janz, Jiri Hron, Przemys■aw Mazur, Katja Hofmann, José Miguel Hernández-Lobato, Sebastian Tschiatschek
Posterior sampling for reinforcement learning (PSRL) is an effective method for balancing exploration and exploitation in reinforcement learning. Randomised value functions (RVF) can be viewed as a promising approach to scaling PSRL. However, we show that most contemporary algorithms combining RVF with neural network function approximation do not possess the properties which make PSRL effective, and provably fail in sparse reward problems. Moreover, we find that propagation of uncertainty, a property of PSRL previously thought important for exploration, does not preclude this failure. We use these insights to design Successor Uncertainties (SU), a cheap and easy to implement RVF algorithm that retains key properties of PSRL. SU is highly effective on hard tabular exploration benchmarks. Furthermore, on the Atari 2600 domain, it surpasses human performance on 38 of 49 games tested (achieving a median human normalised score of 2.09), and outperforms its closest RVF competitor, Bootstrapped DQN, on 36 of those.
****
Trivializations for Gradient-Based Optimization on Manifolds
Mario Lezcano Casado
We introduce a framework to study the transformation of problems with manifold constraints into unconstrained problems through parametrizations in terms of a Euclidean space.
We call these parametrizations trivializations.
We prove conditions under which a trivialization is sound in the context of gradient-based optimization and we show how two large families of trivializations have overall favorable properties, but also suffer from a performance issue.
We then introduce dynamic trivializations, which solve this problem, and we show how these form a family of optimization methods that lie between trivializations and Riemannian gradient descent, and combine the benefits of both of them.
We then show how to implement these two families of trivializations in practice for different matrix manifolds. To this end, we prove a formula for the gradient of the exponential of matrices, which can be of practical interest on its own.
Finally, we show how dynamic trivializations improve the performance of existing methods on standard tasks designed to test long-term memory within neural networks.
****
On the Fairness of Disentangled Representations
Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, Olivier Bachem
Recently there has been a significant interest in learning disentangled representations, as they promise increased interpretability, generalization to unseen scenarios and faster learning on downstream tasks.
In this paper, we investigate the usefulness of different notions of disentanglement for improving the fairness of downstream prediction tasks based on representations.

We consider the setting where the goal is to predict a target variable based on the learned representation of high-dimensional observations (such as images) that depend on both the target variable and an unobserved sensitive variable.
We show that in this setting both the optimal and empirical predictions can be unfair, even if the target variable and the sensitive variable are independent.
Analyzing the representations of more than 12600 trained state-of-the-art disentangled models, we observe that several disentanglement scores are consistently correlated with increased fairness, suggesting that disentanglement may be a useful property to encourage fairness when sensitive variables are not observed.
****
When to use parametric models in reinforcement learning?
Hado P. van Hasselt, Matteo Hessel, John Aslanides
We examine the question of when and how parametric models are most useful in reinforcement learning. In particular, we look at commonalities and differences between parametric models and experience replay. Replay-based learning algorithms share important traits with model-based approaches, including the ability to plan: to use more computation without additional data to improve predictions and behaviour. We discuss when to expect benefits from either approach, and interpret prior work in this context. We hypothesise that, under suitable conditions, replay-based algorithms should be competitive to or better than model-based algorithms if the model is used only to generate fictional transitions from observed states for an update rule that is otherwise model-free. We validated this hypothesis on Atari 2600 video games. The replay-based algorithm attained state-of-the-art data efficiency, improving over prior results with parametric models. Additionally, we discuss different ways to use models. We show that it can be better to plan backward than to plan forward when using models to perform credit assignment (e.g., to directly learn a value or policy), even though the latter seems more common. Finally, we argue and demonstrate that it can be beneficial to plan forward for immediate behaviour, rather than for credit assignment.
****
Ouroboros: On Accelerating Training of Transformer-Based Language Models
Qian Yang, Zhouyuan Huo, Wenlin Wang, Lawrence Carin
Language models are essential for natural language processing (NLP) tasks, such as machine translation and text summarization. Remarkable performance has been demonstrated recently across many NLP domains via a Transformer-based language model with over a billion parameters, verifying the benefits of model size. Model parallelism is required if a model is too large to fit in a single computing device. Current methods for model parallelism either suffer from backward locking in backpropagation or are not applicable to language models. We propose the first model-parallel algorithm that speeds the training of Transformer-based language models. We also prove that our proposed algorithm is guaranteed to converge to critical points for non-convex problems. Extensive experiments on Transformer and Transformer-XL language models demonstrate that the proposed algorithm obtains a much faster speedup beyond data parallelism, with comparable or better accuracy. Code to reproduce experiments is to be found at \url{https://github.com/LaraQianYang/Ouroboros}.
****
MonoForest framework for tree ensemble analysis
Igor Kuralenok, Vasilii Ershov, Igor Labutin
In this work, we introduce a new decision tree ensemble representation framework: instead of using a graph model we transform each tree into a well-known polynomial form. We apply the new representation to three tasks: theoretical analysis, model reduction, and interpretation. The polynomial form of a tree ensemble allows a straightforward interpretation of the original model. In our experiments, it shows comparable results with state-of-the-art interpretation techniques. Another application of the framework is the ensemble-wise pruning: we can drop monomials from the polynomial, based on train data statistics. This way we reduce the model size up to 3 times without loss of its quality. It is possible to show the equivalence of tree shape classes that share the same polynomial. This fact gives us the ability to train a model in one tree's shape and exploit it in anoth

er, which is easier for computation or interpretation. We formulate a problem statement for optimal tree ensemble translation from one form to another and build a greedy solution to this problem.
****

Correlation Priors for Reinforcement Learning
Bastian Alt, Adrian Šoši■, Heinz Koeppl
Many decision-making problems naturally exhibit pronounced structures inherited from the characteristics of the underlying environment. In a Markov decision process
model, for example, two distinct states can have inherently related semantics or encode resembling physical state configurations. This often implies locally correlated transition dynamics among the states. In order to complete a certain task in such environments, the operating agent usually needs to execute a series of temporally and spatially correlated actions. Though there exists a variety of approaches to capture these correlations in continuous state-action domains, a principled solution for discrete environments is missing. In this work, we present a Bayesian learning framework based on Pólya-Gamma augmentation that enables an analogous reasoning in such cases. We demonstrate the framework on a number of common decision-making related problems, such as imitation learning, subgoal extraction, system identification and Bayesian reinforcement learning. By explicitly modeling the underlying correlation structures of these problems, the proposed approach yields superior predictive performance compared to correlation-agnostic models, even when trained on data sets that are an order of magnitude smaller in size.
****

Push-pull Feedback Implements Hierarchical Information Retrieval Efficiently
Xiao Liu, Xiaolong Zou, Zilong Ji, Gengshuo Tian, Yuanyuan Mi, Tiejun Huang, K. Y. Michael Wong, Si Wu
Experimental data has revealed that in addition to feedforward connections, there exist abundant feedback connections in a neural pathway. Although the importance of feedback in neural information processing has been widely recognized in the field, the detailed mechanism of how it works remains largely unknown. Here, we investigate the role of feedback in hierarchical information retrieval. Specifically, we consider a hierarchical network storing the hierarchical categorical information of objects, and information retrieval goes from rough to fine, aided by dynamical push-pull feedback from higher to lower layers. We elucidate that the push (positive) and pull (negative) feedbacks suppress the interferences due to neural correlations between different and the same categories, respectively, and their joint effect improves retrieval performance significantly. Our model agrees with the push-pull phenomenon observed in neural data and sheds light on our understanding of the role of feedback in neural information processing.
****

Calibration tests in multi-class classification: A unifying framework
David Widmann, Fredrik Lindsten, Dave Zachariah
In safety-critical applications a probabilistic model is usually required to be calibrated, i.e., to capture the uncertainty of its predictions accurately. In multi-class classification, calibration of the most confident predictions only is often not sufficient. We propose and study calibration measures for multi-class classification that generalize existing measures such as the expected calibration error, the maximum calibration error, and the maximum mean calibration error. We propose and evaluate empirically different consistent and unbiased estimators for a specific class of measures based on matrix-valued kernels. Importantly, these estimators can be interpreted as test statistics associated with well-defined bounds and approximations of the p-value under the null hypothesis that the model is calibrated, significantly improving the interpretability of calibration measures, which otherwise lack any meaningful unit or scale.
****

Joint Optimization of Tree-based Index and Deep Model for Recommender Systems
Han Zhu, Daqing Chang, Ziru Xu, Pengye Zhang, Xiang Li, Jie He, Han Li, Jian Xu, Kun Gai

Large-scale industrial recommender systems are usually confronted with computational problems due to the enormous corpus size. To retrieve and recommend the most relevant items to users under response time limits, resorting to an efficient index structure is an effective and practical solution. The previous work Tree-based Deep Model (TDM) \cite{zhu2018learning} greatly improves recommendation accuracy using tree index. By indexing items in a tree hierarchy and training a user-node preference prediction model satisfying a max-heap like property in the tree, TDM provides logarithmic computational complexity w.r.t. the corpus size, enabling the use of arbitrary advanced models in candidate retrieval and recommendation.
****
Accurate Uncertainty Estimation and Decomposition in Ensemble Learning
Jeremiah Liu, John Paisley, Marianthi-Anna Kioumourtzoglou, Brent Coull
Ensemble learning is a standard approach to building machine learning systems that capture complex phenomena in real-world data. An important aspect of these systems is the complete and valid quantification of model uncertainty. We introduce a Bayesian nonparametric ensemble (BNE) approach that augments an existing ensemble model to account for different sources of model uncertainty. BNE augments a model’s prediction and distribution functions using Bayesian nonparametric machinery. It has a theoretical guarantee in that it robustly estimates the uncertainty patterns in the data distribution, and can decompose its overall predictive uncertainty into distinct components that are due to different sources of noise and error. We show that our method achieves accurate uncertainty estimates under complex observational noise, and illustrate its real-world utility in terms of uncertainty decomposition and model bias detection for an ensemble in predict air pollution exposures in Eastern Massachusetts, USA.
****
Globally Optimal Learning for Structured Elliptical Losses
Yoav Wald, Nofar Noy, Gal Elidan, Ami Wiesel
Heavy tailed and contaminated data are common in various applications of machine learning. A standard technique to handle regression tasks that involve such data, is to use robust losses, e.g., the popular Huber’s loss.
****
MixMatch: A Holistic Approach to Semi-Supervised Learning
David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, Colin A. Raffel
Semi-supervised learning has proven to be a powerful paradigm for leveraging unlabeled data to mitigate the reliance on large labeled datasets.
In this work, we unify the current dominant approaches for semi-supervised learning to produce a new algorithm, MixMatch, that
guesses low-entropy labels for data-augmented unlabeled examples and mixes labeled and unlabeled data using MixUp.
MixMatch obtains state-of-the-art results by a large margin across many datasets and labeled data amounts. For example,
on CIFAR-10 with 250 labels, we reduce error rate by a factor of 4 (from 38% to 11%) and by a factor of 2 on STL-10.
We also demonstrate how MixMatch can help achieve a dramatically better accuracy-privacy trade-off for differential privacy.
Finally, we perform an ablation study to tease apart which components of MixMatch are most important for its success.
Code is attached.
****
Preventing Gradient Attenuation in Lipschitz Constrained Convolutional Networks
Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B. Grosse, Joern-Henrik Jacobsen
Lipschitz constraints under L2 norm on deep neural networks are useful for provable adversarial robustness bounds, stable training, and Wasserstein distance estimation. While heuristic approaches such as the gradient penalty have seen much practical success, it is challenging to achieve similar practical performance while provably enforcing a Lipschitz constraint. In principle, one can design Lips

chitz constrained architectures using the composition property of Lipschitz func tions, but Anil et al. recently identified a key obstacle to this approach: grad ient norm attenuation. They showed how to circumvent this problem in the case of fully connected networks by designing each layer to be gradient norm preserving . We extend their approach to train scalable, expressive, provably Lipschitz con volutional networks. In particular, we present the Block Convolution Orthogonal Parameterization (BCOP), an expressive parameterization of orthogonal convolutio n operations. We show that even though the space of orthogonal convolutions is d isconnected, the largest connected component of BCOP with 2n channels can repres ent arbitrary BCOP convolutions over n channels. Our BCOP parameterization allow s us to train large convolutional networks with provable Lipschitz bounds. Empir ically, we find that it is competitive with existing approaches to provable adve rsarial robustness and Wasserstein distance estimation.
****

Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder
Ji Feng, Qi-Zhi Cai, Zhi-Hua Zhou
In this work, we consider one challenging training time attack by modifying trai ning data with bounded perturbation, hoping to manipulate the behavior (both tar geted or non-targeted) of any corresponding trained classifier during test time when facing clean samples. To achieve this, we proposed to use an auto-encoder-l ike network to generate such adversarial perturbations on the training data toge ther with one imaginary victim differentiable classifier. The perturbation gener ator will learn to update its weights so as to produce the most harmful noise, a iming to cause the lowest performance for the victim classifier during test time . This can be formulated into a non-linear equality constrained optimization pro blem. Unlike GANs, solving such problem is computationally challenging, we then proposed a simple yet effective procedure to decouple the alternating updates fo r the two networks for stability. By teaching the perturbation generator to hija cking the training trajectory of the victim classifier, the generator can thus l earn to move against the victim classifier step by step. The method proposed in this paper can be easily extended to the label specific setting where the attack er can manipulate the predictions of the victim classifier according to some pre defined rules rather than only making wrong predictions. Experiments on various datasets including CIFAR-10 and a reduced version of ImageNet confirmed the effe ctiveness of the proposed method and empirical results showed that, such bounded perturbations have good transferability across different types of victim classi fiers.
****

Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness
Fanny Yang, Zuowen Wang, Christina Heinze-Deml
This work provides theoretical and empirical evidence that invariance-inducing r egularizers can increase predictive accuracy for worst-case spatial transformati ons (spatial robustness). Evaluated on these adversarially transformed examples, standard and adversarial training with such regularizers achieves a relative er ror reduction of 20% for CIFAR-10 with the same computational budget. This even surpasses handcrafted spatial-equivariant networks. Furthermore, we observe for SVHN, known to have inherent variance in orientation, that robust training also improves standard accuracy on the test set. We prove that this no-trade-off phen omenon holds for adversarial examples from transformation groups.
****

Attentive State-Space Modeling of Disease Progression
Ahmed M. Alaa, Mihaela van der Schaar
Models of disease progression are instrumental for predicting patient outcomes a nd understanding disease dynamics. Existing models provide the patient with prag matic (supervised) predictions of risk, but do not provide the clinician with in telligible (unsupervised) representations of disease pathophysiology. In this pa per, we develop the attentive state-space model, a deep probabilistic model that learns accurate and interpretable structured representations for disease trajec tories. Unlike Markovian state-space models, in which the dynamics are memoryles

s, our model uses an attention mechanism to create "memoryful" dynamics, whereby attention weights determine the dependence of future disease states on past medical history. To learn the model parameters from medical records, we develop an infer ence algorithm that simultaneously learns a compiled inference network and the model parameters, leveraging the attentive state-space representation to construct a "Rao-Blackwellized" variational approximation of the posterior state distribution. Experiments on data from the UK Cystic Fibrosis registry show that our model demonstrates superior predictive accuracy and provides insights into the progression of chronic disease.
****

On two ways to use determinantal point processes for Monte Carlo integration
Guillaume Gautier, Rémi Bardenet, Michal Valko
When approximating an integral by a weighted sum of function evaluations, determinantal point processes (DPPs) provide a way to enforce repulsion between the evaluation points.
This negative dependence is encoded by a kernel.
Fifteen years before the discovery of DPPs, Ermakov & Zolotukhin (EZ, 1960) had the intuition of sampling a DPP and solving a linear system to compute an unbiased Monte Carlo estimator of the integral.
In the absence of DPP machinery to derive an efficient sampler and analyze their estimator, the idea of Monte Carlo integration with DPPs was stored in the cellar of numerical integration.
Recently, Bardenet & Hardy (BH, 2019) came up with a more natural estimator with a fast central limit theorem (CLT).
In this paper, we first take the EZ estimator out of the cellar, and analyze it using modern arguments.
Second, we provide an efficient implementation to sample exactly a particular multidimensional DPP called multivariate Jacobi ensemble.
The latter satisfies the assumptions of the aforementioned CLT.
Third, our new implementation lets us investigate the behavior of the two unbiased Monte Carlo estimators in yet unexplored regimes.
We demonstrate experimentally good properties when the kernel is adapted to basis of functions in which the integrand is sparse or has fast-decaying coefficients.
If such a basis and the level of sparsity are known (e.g., we integrate a linear combination of kernel eigenfunctions), the EZ estimator can be the right choice, but otherwise it can display an erratic behavior.
****

ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls
Jinjin Tian, Aaditya Ramdas
Major internet companies routinely perform tens of thousands of A/B tests each year. Such large-scale sequential experimentation has resulted in a recent spurt of new algorithms that can provably control the false discovery rate (FDR) in a fully online fashion. However, current state-of-the-art adaptive algorithms can suffer from a significant loss in power if null p-values are conservative (stochastically larger than the uniform distribution), a situation that occurs frequently in practice. In this work, we introduce a new adaptive discarding method called ADDIS that provably controls the FDR and achieves the best of both worlds: it enjoys appreciable power increase over all existing methods if nulls are conservative (the practical case), and rarely loses power if nulls are exactly uniformly distributed (the ideal case). We provide several practical insights on robust choices of tuning parameters, and extend the idea to asynchronous and offline settings as well.
****

Controllable Text-to-Image Generation
Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, Philip Torr
In this paper, we propose a novel controllable text-to-image generative adversarial network (ControlGAN), which can effectively synthesise high-quality images and also control parts of the image generation according to natural language desc

riptions. To achieve this, we introduce a word-level spatial and channel-wise attention-driven generator that can disentangle different visual attributes, and allow the model to focus on generating and manipulating subregions corresponding to the most relevant words. Also, a word-level discriminator is proposed to provide fine-grained supervisory feedback by correlating words with image regions, facilitating training an effective generator which is able to manipulate specific visual attributes without affecting the generation of other content. Furthermore, perceptual loss is adopted to reduce the randomness involved in the image generation, and to encourage the generator to manipulate specific attributes required in the modified text. Extensive experiments on benchmark datasets demonstrate that our method outperforms existing state of the art, and is able to effectively manipulate synthetic images using natural language descriptions.
****

Exploring Algorithmic Fairness in Robust Graph Covering Problems

Aida Rahmattalabi, Phebe Vayanos, Anthony Fulginiti, Eric Rice, Bryan Wilder, Amulya Yadav, Milind Tambe

Fueled by algorithmic advances, AI algorithms are increasingly being deployed in settings subject to unanticipated challenges with complex social effects. Motivated by real-world deployment of AI driven, social-network based suicide prevention and landslide risk management interventions, this paper focuses on a robust graph covering problem subject to group fairness constraints. We show that, in the absence of fairness constraints, state-of-the-art algorithms for the robust graph covering problem result in biased node coverage: they tend to discriminate individuals (nodes) based on membership in traditionally marginalized groups. To remediate this issue, we propose a novel formulation of the robust covering problem with fairness constraints and a tractable approximation scheme applicable to real world instances. We provide a formal analysis of the price of group fairness (PoF) for this problem, where we show that uncertainty can lead to greater PoF. We demonstrate the effectiveness of our approach on several real-world social networks. Our method yields competitive node coverage while significantly improving group fairness relative to state-of-the-art methods.
****

Fast Convergence of Natural Gradient Descent for Over-Parameterized Neural Networks

Guodong Zhang, James Martens, Roger B. Grosse

Natural gradient descent 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, especially for \emph{non-linear} networks. In this work, we analyze for the first time the speed of convergence to global optimum for natural gradient descent on non-linear neural networks with the squared error loss. We identify two conditions which guarantee the global convergence: (1) the Jacobian matrix (of network's output for all training cases w.r.t the parameters) is full row rank and (2) the Jacobian matrix is stable for small perturbations around the initialization. For two-layer ReLU neural networks (i.e. with one hidden layer), we prove that these two conditions do hold throughout the training under the assumptions that the inputs do not degenerate and the network is over-parameterized. We further extend our analysis to more general loss function with similar convergence property. Lastly, we show that K-FAC, an approximate natural gradient descent method, also converges to global minima under the same assumptions.
****

Reducing the variance in online optimization by transporting past gradients

Sébastien Arnold, Pierre-Antoine Manzagol, Reza Babanezhad Harikandeh, Ioannis Mitliagkas, Nicolas Le Roux

Most stochastic optimization methods use gradients once before discarding them. While variance reduction methods have shown that reusing past gradients can be beneficial when there is a finite number of datapoints, they do not easily extend to the online setting. One issue is the staleness due to using past gradients. We propose to correct this staleness using the idea of {\em implicit gradient transport} (IGT) which transforms gradients computed at previous iterates into gra

dients evaluated at the current iterate without using the Hessian explicitly. In addition to reducing the variance and bias of our updates over time, IGT can be used as a drop-in replacement for the gradient estimate in a number of well-understood methods such as heavy ball or Adam. We show experimentally that it achieves state-of-the-art results on a wide range of architectures and benchmarks. Additionally, the IGT gradient estimator yields the optimal asymptotic convergence rate for online stochastic optimization in the restricted setting where the Hessians of all component functions are equal.
****
Deep Multi-State Dynamic Recurrent Neural Networks Operating on Wavelet Based Neural Features for Robust Brain Machine Interfaces
Benyamin Allahgholizadeh Haghi, Spencer Kellis, Sahil Shah, Maitreyi Ashok, Luke Bashford, Daniel Kramer, Brian Lee, Charles Liu, Richard Andersen, Azita Emami
We present a new deep multi-state Dynamic Recurrent Neural Network (DRNN) architecture for Brain Machine Interface (BMI) applications. Our DRNN is used to predict Cartesian representation of a computer cursor movement kinematics from open-loop neural data recorded from the posterior parietal cortex (PPC) of a human subject in a BMI system. We design the algorithm to achieve a reasonable trade-off between performance and robustness, and we constrain memory usage in favor of future hardware implementation. We feed the predictions of the network back to the input to improve prediction performance and robustness. We apply a scheduled sampling approach to the model in order to solve a statistical distribution mismatch between the ground truth and predictions. Additionally, we configure a small DRNN to operate with a short history of input, reducing the required buffering of input data and number of memory accesses. This configuration lowers the expected power consumption in a neural network accelerator. Operating on wavelet-based neural features, we show that the average performance of DRNN surpasses other state-of-the-art methods in the literature on both single- and multi-day data recorded over 43 days. Results show that multi-state DRNN has the potential to model the nonlinear relationships between the neural data and kinematics for robust BMIs.
****
Graph Normalizing Flows
Jenny Liu, Aviral Kumar, Jimmy Ba, Jamie Kiros, Kevin Swersky
We introduce graph normalizing flows: a new, reversible graph neural network model for prediction and generation. On supervised tasks, graph normalizing flows perform similarly to message passing neural networks, but at a significantly reduced memory footprint, allowing them to scale to larger graphs. In the unsupervised case, we combine graph normalizing flows with a novel graph auto-encoder to create a generative model of graph structures. Our model is permutation-invariant, generating entire graphs with a single feed-forward pass, and achieves competitive results with the state-of-the art auto-regressive models, while being better suited to parallel computing architectures.
****
Cascaded Dilated Dense Network with Two-step Data Consistency for MRI Reconstruction
Hao Zheng, Faming Fang, Guixu Zhang
Compressed Sensing MRI (CS-MRI) aims at reconstrcuting de-aliased images from sub-Nyquist sampling k-space data to accelerate MR Imaging. Inspired by recent deep learning methods, we propose a Cascaded Dilated Dense Network (CDDN) for MRI reconstruction. Dense blocks with residual connection are used to restore clear images step by step and dilated convolution is introduced for expanding receptive field without taking more network parameters. After each sub-network, we use a novel two-step Data Consistency (DC) operation in k-space. We convert the complex result from first DC operation to real-valued images and applied another sampled \emph{k}-space data replacement. Extensive experiments demonstrate that the proposed CDDN with two-step DC achieves state-of-art result.
****
Neural networks grown and self-organized by noise
Guruprasad Raghavan, Matt Thomson

Living neural networks emerge through a process of growth and self-organization
that begins with a single cell and results in a brain, an organized and function
al computational device. Artificial neural networks, however, rely on human-desi
gned, hand-programmed architectures for their remarkable performance. Can we dev
elop artificial computational devices that can grow and self-organize without hu
man intervention? In this paper, we propose a biologically inspired developmenta
l algorithm that can ‘grow’ a functional, layered neural network from a single i
nitial cell. The algorithm organizes inter-layer connections to construct retino
topic pooling layers. Our approach is inspired by the mechanisms employed by the
early visual system to wire the retina to the lateral geniculate nucleus (LGN),
days before animals open their eyes. The key ingredients for robust self-organ
ization are an emergent spontaneous spatiotemporal activity wave in the first la
yer and a local learning rule in the second layer that ‘learns’ the underlying a
ctivity pattern in the first layer. The algorithm is adaptable to a wide-range
of input-layer geometries, robust to malfunctioning units in the first layer, an
d so can be used to successfully grow and self-organize pooling architectures of
different pool-sizes and shapes. The algorithm provides a primitive procedure f
or constructing layered neural networks through growth and self-organization. We
also demonstrate that networks grown from a single unit perform as well as hand
-crafted networks on MNIST. Broadly, our work shows that biologically inspired d
evelopmental algorithms can be applied to autonomously grow functional `brains'
in-silico.
****
Likelihood Ratios for Out-of-Distribution Detection
Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, J
oshua Dillon, Balaji Lakshminarayanan
Discriminative neural networks offer little or no performance guarantees when de
ployed on data not generated by the same process as the training distribution. O
n such out-of-distribution (OOD) inputs, the prediction may not only be erroneou
s, but confidently so, limiting the safe deployment of classifiers in real-world
applications. One such challenging application is bacteria identification based
on genomic sequences, which holds the promise of early detection of diseases, b
ut requires a model that can output low confidence predictions on OOD genomic se
quences from new bacteria that were not present in the training data. We introdu
ce a genomics dataset for OOD detection that allows other researchers to benchma
rk progress on this important problem. We investigate deep generative model base
d approaches for OOD detection and observe that the likelihood score is heavily
affected by population level background statistics. We propose a likelihood rati
o method for deep generative models which effectively corrects for these confoun
ding background statistics. We benchmark the OOD detection performance of the pr
oposed method against existing approaches on the genomics dataset and show that
our method achieves state-of-the-art performance. Finally, we demonstrate the ge
nerality of the proposed method by showing that it significantly improves OOD de
tection when applied to deep generative models of images.
****
Root Mean Square Layer Normalization
Biao Zhang, Rico Sennrich
Layer normalization (LayerNorm) has been successfully applied to various deep ne
ural networks to help stabilize training and boost model convergence because of
its capability in handling re-centering and re-scaling of both inputs and weight
matrix. However, the computational overhead introduced by LayerNorm makes these
improvements expensive and significantly slows the underlying network, e.g. RNN
in particular. In this paper, we hypothesize that re-centering invariance in La
yerNorm is dispensable and propose root mean square layer normalization, or RMSN
orm. RMSNorm regularizes the summed inputs to a neuron in one layer according to
root mean square (RMS), giving the model re-scaling invariance property and imp
licit learning rate adaptation ability. RMSNorm is computationally simpler and t
hus more efficient than LayerNorm. We also present partial RMSNorm, or pRMSNorm
where the RMS is estimated from p% of the summed inputs without breaking the abo
ve properties. Extensive experiments on several tasks using diverse network arch

itectures show that RMSNorm achieves comparable performance against LayerNorm but reduces the running time by 7%~64% on different models. Source code is available at https://github.com/bzhangGo/rmsnorm.
****
HyperGCN: A New Method For Training Graph Convolutional Networks on Hypergraphs
Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, Partha Talukdar
In many real-world network datasets such as co-authorship, co-citation, email communication, etc., relationships are complex and go beyond pairwise. Hypergraphs provide a flexible and natural modeling tool to model such complex relationships. The obvious existence of such complex relationships in many real-world networks naturaly motivates the problem of learning with hypergraphs. A popular learning paradigm is hypergraph-based semi-supervised learning (SSL) where the goal is to assign labels to initially unlabeled vertices in a hypergraph. Motivated by the fact that a graph convolutional network (GCN) has been effective for graph-based SSL, we propose HyperGCN, a novel GCN for SSL on attributed hypergraphs. Additionally, we show how HyperGCN can be used as a learning-based approach for combinatorial optimisation on NP-hard hypergraph problems. We demonstrate HyperGCN's effectiveness through detailed experimentation on real-world hypergraphs. We have made HyperGCN's source code available to foster reproducible research.
****
Asymptotics for Sketching in Least Squares Regression
Edgar Dobriban, Sifan Liu
We consider a least squares regression problem where the data has been generated from a linear model, and we are interested to learn the unknown regression parameters. We consider "sketch-and-solve" methods that randomly project the data first, and do regression after. Previous works have analyzed the statistical and computational performance of such methods. However, the existing analysis is not fine-grained enough to show the fundamental differences between various methods, such as the Subsampled Randomized Hadamard Transform (SRHT) and Gaussian projections. In this paper, we make progress on this problem, working in an asymptotic framework where the number of datapoints and dimension of features goes to infinity. We find the limits of the accuracy loss (for estimation and test error) incurred by popular sketching methods. We show separation between different methods, so that SRHT is better than Gaussian projections. Our theoretical results are verified on both real and synthetic data. The analysis of SRHT relies on novel methods from random matrix theory that may be of independent interest.
****
Gradient Dynamics of Shallow Univariate ReLU Networks
Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, Joan Bruna
We present a theoretical and empirical study of the gradient dynamics of overparameterized shallow ReLU networks with one-dimensional input, solving least-squares interpolation. We show that the gradient dynamics of such networks are determined by the gradient flow in a non-redundant parameterization of the network function. We examine the principal qualitative features of this gradient flow. In particular, we determine conditions for two learning regimes: \emph{kernel} and \emph{adaptive}, which depend both on the relative magnitude of initialization of weights in different layers and the asymptotic behavior of initialization coefficients in the limit of large network widths. We show that learning in the kernel regime yields smooth interpolants, minimizing curvature, and reduces to \emph{cubic splines} for uniform initializations. Learning in the adaptive regime favors instead \emph{linear splines}, where knots cluster adaptively at the sample points.
****
Chirality Nets for Human Pose Regression
Raymond Yeh, Yuan-Ting Hu, Alexander Schwing
We propose Chirality Nets, a family of deep nets that is equivariant to the “chirality transform,” i.e., the transformation to create a chiral pair. Through parameter sharing, odd and even symmetry, we propose and prove variants of standard

building blocks of deep nets that satisfy the equivariance property, including fully connected layers, convolutional layers, batch-normalization, and LSTM/GRU cells. The proposed layers lead to a more data efficient representation and a reduction in computation by exploiting symmetry. We evaluate chirality nets on the task of human pose regression, which naturally exploits the left/right mirroring of the human body. We study three pose regression tasks: 3D pose estimation from video, 2D pose forecasting, and skeleton based activity recognition. Our approach achieves/matches state-of-the-art results, with more significant gains on small datasets and limited-data settings.
****
TAB-VCR: Tags and Attributes based VCR Baselines
Jingxiang Lin, Unnat Jain, Alexander Schwing
Reasoning is an important ability that we learn from a very early age. Yet, reasoning is extremely hard for algorithms. Despite impressive recent progress that has been reported on tasks that necessitate reasoning, such as visual question answering and visual dialog, models often exploit biases in datasets. To develop models with better reasoning abilities, recently, the new visual commonsense reasoning(VCR) task has been introduced. Not only do models have to answer questions, but also do they have to provide a reason for the given answer. The proposed baseline achieved compelling results, leveraging a meticulously designed model composed of LSTM modules and attention nets. Here we show that a much simpler model obtained by ablating and pruning the existing intricate baseline can perform better with half the number of trainable parameters. By associating visual features with attribute information and better text to image grounding, we obtain further improvements for our simpler & effective baseline, TAB-VCR. We show that this approach results in a 5.3%, 4.4% and 6.5% absolute improvement over the previous state-of-the-art on question answering, answer justification and holistic VCR. Webpage: https://deanplayerljx.github.io/tabvcr/
****
Multiclass Performance Metric Elicitation
Gaurush Hiranandani, Shant Boodaghians, Ruta Mehta, Oluwasanmi O. Koyejo
Metric Elicitation is a principled framework for selecting the performance metric that best reflects implicit user preferences. However, available strategies have so far been limited to binary classification. In this paper, we propose novel strategies for eliciting multiclass classification performance metrics using only relative preference feedback. We also show that the strategies are robust to both finite sample and feedback noise.
****
Assessing Social and Intersectional Biases in Contextualized Word Representations
Yi Chern Tan, L. Elisa Celis
Social bias in machine learning has drawn significant attention, with work ranging from demonstrations of bias in a multitude of applications, curating definitions of fairness for different contexts, to developing algorithms to mitigate bias. In natural language processing, gender bias has been shown to exist in context-free word embeddings. Recently, contextual word representations have outperformed word embeddings in several downstream NLP tasks. These word representations are conditioned on their context within a sentence, and can also be used to encode the entire sentence. In this paper, we analyze the extent to which state-of-the-art models for contextual word representations, such as BERT and GPT-2, encode biases with respect to gender, race, and intersectional identities. Towards this, we propose assessing bias at the contextual word level. This novel approach captures the contextual effects of bias missing in context-free word embeddings, yet avoids confounding effects that underestimate bias at the sentence encoding level. We demonstrate evidence of bias at the corpus level, find varying evidence of bias in embedding association tests, show in particular that racial bias is strongly encoded in contextual word models, and observe that bias effects for intersectional minorities are exacerbated beyond their constituent minority identities. Further, evaluating bias effects at the contextual word level captures biases that are not captured at the sentence level, confirming the need for our n

ovel approach.
****
Likelihood-Free Overcomplete ICA and Applications In Causal Discovery
Chenwei DING, Mingming Gong, Kun Zhang, Dacheng Tao
Causal discovery witnessed significant progress over the past decades. In particular, many recent causal discovery methods make use of independent, non-Gaussian noise to achieve identifiability of the causal models. Existence of hidden direct common causes, or confounders, generally makes causal discovery more difficult; whenever they are present, the corresponding causal discovery algorithms can be seen as extensions of overcomplete independent component analysis (OICA). However, existing OICA algorithms usually make strong parametric assumptions on the distribution of independent components, which may be violated on real data, leading to sub-optimal or even wrong solutions. In addition, existing OICA algorithms rely on the Expectation Maximization (EM) procedure that requires computationally expensive inference of the posterior distribution of independent components. To tackle these problems, we present a Likelihood-Free Overcomplete ICA algorithm (LFOICA) that estimates the mixing matrix directly by back-propagation without any explicit assumptions on the density function of independent components. Thanks to its computational efficiency, the proposed method makes a number of causal discovery procedures much more practically feasible. For illustrative purposes, we demonstrate the computational efficiency and efficacy of our method in two causal discovery tasks on both synthetic and real data.
****
MaCow: Masked Convolutional Generative Flow
Xuezhe Ma, Xiang Kong, Shanghang Zhang, Eduard Hovy
Flow-based generative models, conceptually attractive due to tractability of both the exact log-likelihood computation and latent-variable inference, and efficiency of both training and sampling, has led to a number of impressive empirical successes and spawned many advanced variants and theoretical investigations.
Despite their computational efficiency, the density estimation performance of flow-based generative models significantly falls behind those of state-of-the-art autoregressive models. In this work, we introduce masked convolutional generative flow (MaCow), a simple yet effective architecture of generative flow using masked convolution. By restricting the local connectivity in a small kernel, MaCow enjoys the properties of fast and stable training, and efficient sampling, while achieving significant improvements over Glow for density estimation on standard image benchmarks, considerably narrowing the gap to autoregressive models.
****
Batched Multi-armed Bandits Problem
Zijun Gao, Yanjun Han, Zhimei Ren, Zhengqing Zhou
In this paper, we study the multi-armed bandit problem in the batched setting where the employed policy must split data into a small number of batches. While the minimax regret for the two-armed stochastic bandits has been completely characterized in \cite{perchet2016batched}, the effect of the number of arms on the regret for the multi-armed case is still open. Moreover, the question whether adaptively chosen batch sizes will help to reduce the regret also remains underexplored. In this paper, we propose the BaSE (batched successive elimination) policy to achieve the rate-optimal regrets (within logarithmic factors) for batched multi-armed bandits, with matching lower bounds even if the batch sizes are determined in an adaptive manner.
****
High-Quality Self-Supervised Deep Image Denoising
Samuli Laine, Tero Karras, Jaakko Lehtinen, Timo Aila
We describe a novel method for training high-quality image denoising models based on unorganized collections of corrupted images. The training does not need access to clean reference images, or explicit pairs of corrupted images, and can thus be applied in situations where such data is unacceptably expensive or impossible to acquire. We build on a recent technique that removes the need for reference data by employing networks with a "blind spot" in the receptive field, and significantly improve two key aspects: image quality and training efficiency. Our

result quality is on par with state-of-the-art neural network denoisers in the case of i.i.d. additive Gaussian noise, and not far behind with Poisson and impulse noise. We also successfully handle cases where parameters of the noise model are variable and/or unknown in both training and evaluation data.
****

Generalization in multitask deep neural classifiers: a statistical physics approach

Anthony Ndirango, Tyler Lee

A proper understanding of the striking generalization abilities of deep neural networks presents an enduring puzzle. Recently, there has been a growing body of numerically-grounded theoretical work that has contributed important insights to the theory of learning in deep neural nets. There has also been a recent interest in extending these analyses to understanding how multitask learning can further improve the generalization capacity of deep neural nets. These studies deal almost exclusively with regression tasks which are amenable to existing analytical techniques. We develop an analytic theory of the nonlinear dynamics of generalization of deep neural networks trained to solve classification tasks using softmax outputs and cross-entropy loss, addressing both single task and multitask settings. We do so by adapting techniques from the statistical physics of disordered systems, accounting for both finite size datasets and correlated outputs induced by the training dynamics. We discuss the validity of our theoretical results in comparison to a comprehensive suite of numerical experiments. Our analysis provides theoretical support for the intuition that the performance of multitask learning is determined by the noisiness of the tasks and how well their input features align with each other. Highly related, clean tasks benefit each other, whereas unrelated, clean tasks can be detrimental to individual task performance.
****

Causal Regularization

Dominik Janzing

We argue that regularizing terms in standard regression methods not only help against overfitting finite data, but sometimes also help in getting better causal models. We first consider a multi-dimensional variable linearly influencing a target variable with some multi-dimensional unobserved common cause, where the confounding effect can be decreased by keeping the penalizing term in Ridge and Lasso regression even in the population limit. The reason is a close analogy between overfitting and confounding observed for our toy model. In the case of overfitting, we can choose regularization constants via cross validation, but here we choose the regularization constant by first estimating the strength of confounding, which yielded reasonable results for simulated and real data. Further, we show a ‘causal generalization bound’ which states (subject to our particular model of confounding) that the error made by interpreting any non-linear regression as causal model can be bounded from above whenever functions are taken from a not too rich class.
****

Locality-Sensitive Hashing for f-Divergences: Mutual Information Loss and Beyond

Lin Chen, Hossein Esfandiari, Gang Fu, Vahab Mirrokni

Computing approximate nearest neighbors in high dimensional spaces is a central problem in large-scale data mining with a wide range of applications in machine learning and data science. A popular and effective technique in computing nearest neighbors approximately is the locality-sensitive hashing (LSH) scheme. In this paper, we aim to develop LSH schemes for distance functions that measure the distance between two probability distributions, particularly for f-divergences as well as a generalization to capture mutual information loss. First, we provide a general framework to design LHS schemes for f-divergence distance functions and develop LSH schemes for the generalized Jensen-Shannon divergence and triangular discrimination in this framework. We show a two-sided approximation result for approximation of the generalized Jensen-Shannon divergence by the Hellinger distance, which may be of independent interest. Next, we show a general method of reducing the problem of designing an LSH scheme for a Krein kernel (which can be expressed as the difference of two positive definite kernels) to the problem of

maximum inner product search. We exemplify this method by applying it to the mutual information loss, due to its several important applications such as model compression.
****
Augmented Neural ODEs
Emilien Dupont, Arnaud Doucet, Yee Whye Teh
We show that Neural Ordinary Differential Equations (ODEs) learn representations that preserve the topology of the input space and prove that this implies the existence of functions Neural ODEs cannot represent. To address these limitations, we introduce Augmented Neural ODEs which, in addition to being more expressive models, are empirically more stable, generalize better and have a lower computational cost than Neural ODEs.
****
Efficient Smooth Non-Convex Stochastic Compositional Optimization via Stochastic Recursive Gradient Descent
Wenqing Hu, Chris Junchi Li, Xiangru Lian, Ji Liu, Huizhuo Yuan
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.
****
Regularizing Trajectory Optimization with Denoising Autoencoders
Rinu Boney, Norman Di Palo, Mathias Berglund, Alexander Ilin, Juho Kannala, Antti Rasmus, Harri Valpola
Trajectory optimization using a learned model of the environment is one of the core elements of model-based reinforcement learning. This procedure often suffers from exploiting inaccuracies of the learned model. We propose to regularize trajectory optimization by means of a denoising autoencoder that is trained on the same trajectories as the model of the environment. We show that the proposed regularization leads to improved planning with both gradient-based and gradient-free optimizers. We also demonstrate that using regularized trajectory optimization leads to rapid initial learning in a set of popular motor control tasks, which suggests that the proposed approach can be a useful tool for improving sample efficiency.
****
Multi-Criteria Dimensionality Reduction with Applications to Fairness
Uthaipon Tantipongpipat, Samira Samadi, Mohit Singh, Jamie H. Morgenstern, Santosh Vempala
Dimensionality reduction is a classical technique widely used for data analysis. One foundational instantiation is Principal Component Analysis (PCA),
which minimizes the average reconstruction error. In this paper, we introduce the multi-criteria dimensionality reduction problem where we are given multiple objectives that need to be optimized simultaneously. As an application, our model captures several fairness criteria for dimensionality reduction such as the Fair-PCA problem introduced by Samadi et al. [NeurIPS18] and the Nash Social Welfare (NSW) problem. In the Fair-PCA problem, the input data is divided into k groups, and the goal is to find a single d-dimensional representation for all groups for which the maximum reconstruction error of any one group is minimized. In NSW the goal is to maximize the product of the individual variances of the groups achieved by the common low-dimensinal space.
****
Structured and Deep Similarity Matching via Structured and Deep Hebbian Networks
Dina Obeid, Hugo Ramambason, Cengiz Pehlevan
Synaptic plasticity is widely accepted to be the mechanism behind learning in the brain's neural networks. A central question is how synapses, with access to only local information about the network, can still organize collectively and perform circuit-wide learning in an efficient manner. In single-layered and all-to-all connected neural networks, local plasticity has been shown to implement gradient-based learning on a class of cost functions that contain a term that aligns

the similarity of outputs to the similarity of inputs. Whether such cost functions exist for networks with other architectures is not known. In this paper, we introduce structured and deep similarity matching cost functions, and show how they can be optimized in a gradient-based manner by neural networks with local learning rules. These networks extend F\"oldiak’s Hebbian/Anti-Hebbian network to deep architectures and structured feedforward, lateral and feedback connections. Credit assignment problem is solved elegantly by a factorization of the dual learning objective to synapse specific local objectives. Simulations show that our networks learn meaningful features.
****
Neural Trust Region/Proximal Policy Optimization Attains Globally Optimal Policy
Boyi Liu, Qi Cai, Zhuoran Yang, Zhaoran Wang
Proximal policy optimization and trust region policy optimization (PPO and TRPO) with actor and critic parametrized by neural networks achieve significant empirical success in deep reinforcement learning. However, due to nonconvexity, the global convergence of PPO and TRPO remains less understood, which separates theory from practice. In this paper, we prove that a variant of PPO and TRPO equipped with overparametrized neural networks converges to the globally optimal policy at a sublinear rate. The key to our analysis is the global convergence of infinite-dimensional mirror descent under a notion of one-point monotonicity, where the gradient and iterate are instantiated by neural networks. In particular, the desirable representation power and optimization geometry induced by the overparametrization of such neural networks allow them to accurately approximate the infinite-dimensional gradient and iterate.
****
ANODEV2: A Coupled Neural ODE Framework
Tianjun Zhang, Zhewei Yao, Amir Gholami, Joseph E. Gonzalez, Kurt Keutzer, Michael W. Mahoney, George Biros
It has been observed that residual networks can be viewed as the explicit Euler discretization of an Ordinary Differential Equation (ODE). This observation motivated the introduction of so-called Neural ODEs, in which other discretization schemes and/or adaptive time stepping techniques can be used to improve the performance of residual networks. Here, we propose \OURS, which extends this approach by introducing a framework that allows ODE-based evolution for both the weights and the activations, in a coupled formulation. Such an approach provides more modeling flexibility, and it can help with generalization performance. We present the formulation of \OURS, derive optimality conditions, and implement the coupled framework in PyTorch. We present empirical results using several different configurations of \OURS, testing them on the CIFAR-10 dataset. We report results showing that our coupled ODE-based framework is indeed trainable, and that it achieves higher accuracy, compared to the baseline ResNet network and the recently-proposed Neural ODE approach.
****
Learning Neural Networks with Adaptive Regularization
Han Zhao, Yao-Hung Hubert Tsai, Russ R. Salakhutdinov, Geoffrey J. Gordon
Feed-forward neural networks can be understood as a combination of an intermediate representation and a linear hypothesis. While most previous works aim to diversify the representations, we explore the complementary direction by performing an adaptive and data-dependent regularization motivated by the empirical Bayes method. Specifically, we propose to construct a matrix-variate normal prior (on weights) whose covariance matrix has a Kronecker product structure. This structure is designed to capture the correlations in neurons through backpropagation. Under the assumption of this Kronecker factorization, the prior encourages neurons to borrow statistical strength from one another. Hence, it leads to an adaptive and data-dependent regularization when training networks on small datasets. To optimize the model, we present an efficient block coordinate descent algorithm with analytical solutions. Empirically, we demonstrate that the proposed method helps networks converge to local optima with smaller stable ranks and spectral norms. These properties suggest better generalizations and we present empirical results to support this expectation. We also verify the effectiveness of the appro

ach on multiclass classification and multitask regression problems with various network structures. Our code is publicly available at:~\url{https://github.com/yaohungt/Adaptive-Regularization-Neural-Network}.
****
Turbo Autoencoder: Deep learning based channel codes for point-to-point communication channels
Yihan Jiang, Hyeji Kim, Himanshu Asnani, Sreeram Kannan, Sewoong Oh, Pramod Viswanath
Designing codes that combat the noise in a communication medium has remained a significant area of research in information theory as well as wireless communications. Asymptotically optimal channel codes have been developed by mathematicians for communicating under canonical models after over 60 years of research. On the other hand, in many non-canonical channel settings, optimal codes do not exist and the codes designed for canonical models are adapted via heuristics to these channels and are thus not guaranteed to be optimal. In this work, we make significant progress on this problem by designing a fully end-to-end jointly trained neural encoder and decoder, namely, Turbo Autoencoder (TurboAE), with the following contributions: (a) under moderate block lengths, TurboAE approaches state-of-the-art performance under canonical channels; (b) moreover, TurboAE outperforms the state-of-the-art codes under non-canonical settings in terms of reliability. TurboAE shows that the development of channel coding design can be automated via deep learning, with near-optimal performance.
****
DetNAS: Backbone Search for Object Detection
Yukang Chen, Tong Yang, Xiangyu Zhang, GAOFENG MENG, Xinyu Xiao, Jian Sun
Object detectors are usually equipped with backbone networks designed for image classification. It might be sub-optimal because of the gap between the tasks of image classification and object detection. In this work, we present DetNAS to use Neural Architecture Search (NAS) for the design of better backbones for object detection. It is non-trivial because detection training typically needs ImageNetpre-training while NAS systems require accuracies on the target detection task as supervisory signals. Based on the technique of one-shot supernet, which contains all possible networks in the search space, we propose a framework for backbone search on object detection. We train the supernet under the typical detector training schedule: ImageNet pre-training and detection fine-tuning. Then, the architecture search is performed on the trained supernet, using the detection task as the guidance. This framework makes NAS on backbones very efficient. In experiments, we show the effectiveness of DetNAS on various detectors, for instance, one-stage RetinaNetand the two-stage FPN. We empirically find that networks searched on object detection shows consistent superiority compared to those searched on ImageNet classification. The resulting architecture achieves superior performance than hand-crafted networks on COCO with much less FLOPs complexity.
****
Nonlinear scaling of resource allocation in sensory bottlenecks
Laura Rose Edmondson, Alejandro Jimenez Rodriguez, Hannes P. Saal
In many sensory systems, information transmission is constrained by a bottleneck,
where the number of output neurons is vastly smaller than the number of input
neurons. Efficient coding theory predicts that in these scenarios the brain should
allocate its limited resources by removing redundant information. Previous work
has typically assumed that receptors are uniformly distributed across the sensory
sheet, when in reality these vary in density, often by an order of magnitude. How,
then, should the brain efficiently allocate output neurons when the density of input
neurons is nonuniform? Here, we show analytically and numerically that resource
allocation scales nonlinearly in efficient coding models that maximize information

transfer, when inputs arise from separate regions with different receptor densities.
Importantly, the proportion of output neurons allocated to a given input region changes depending on the width of the bottleneck, and thus cannot be predicted from input density or region size alone. Narrow bottlenecks favor magnification of
high density input regions, while wider bottlenecks often cause contraction. Our results demonstrate that both expansion and contraction of sensory input regions can arise in efficient coding models and that the final allocation crucially depends
on the neural resources made available.
****
What the Vec? Towards Probabilistically Grounded Embeddings
Carl Allen, Ivana Balazevic, Timothy Hospedales
Word2Vec (W2V) and Glove are popular word embedding algorithms that perform well on a variety of natural language processing tasks. The algorithms are fast, efficient and their embeddings widely used. Moreover, the W2V algorithm has recently been adopted in the field of graph embedding, where it underpins several leading algorithms. However, despite their ubiquity and the relative simplicity of their common architecture, what the embedding parameters of W2V and Glove learn, and why that it useful in downstream tasks largely remains a mystery. We show that different interactions of PMI vectors encode semantic properties that can be captured in low dimensional word embeddings by suitable projection, theoretically explaining why the embeddings of W2V and Glove work, and, in turn, revealing an interesting mathematical interconnection between the semantic relationships of relatedness, similarity, paraphrase and analogy.
****
Diffusion Improves Graph Learning
Johannes Gasteiger, Stefan Weißenberger, Stephan Günnemann
Graph convolution is the core of most Graph Neural Networks (GNNs) and usually approximated by message passing between direct (one-hop) neighbors. In this work, we remove the restriction of using only the direct neighbors by introducing a powerful, yet spatially localized graph convolution: Graph diffusion convolution (GDC). GDC leverages generalized graph diffusion, examples of which are the heat kernel and personalized PageRank. It alleviates the problem of noisy and often arbitrarily defined edges in real graphs. We show that GDC is closely related to spectral-based models and thus combines the strengths of both spatial (message passing) and spectral methods. We demonstrate that replacing message passing with graph diffusion convolution consistently leads to significant performance improvements across a wide range of models on both supervised and unsupervised tasks and a variety of datasets. Furthermore, GDC is not limited to GNNs but can trivially be combined with any graph-based model or algorithm (e.g. spectral clustering) without requiring any changes to the latter or affecting its computational complexity. Our implementation is available online.
****
Inverting Deep Generative models, One layer at a time
Qi Lei, Ajil Jalal, Inderjit S. Dhillon, Alexandros G. Dimakis
We study the problem of inverting a deep generative model with ReLU activations.

Inversion corresponds to finding a latent code vector that explains observed measurements as much as possible.
In most prior works this is performed by attempting to solve a non-convex optimization problem involving the generator.
In this paper we obtain several novel theoretical results for the inversion problem.
****
Sample Complexity of Learning Mixture of Sparse Linear Regressions
Akshay Krishnamurthy, Arya Mazumdar, Andrew McGregor, Soumyabrata Pal
In the problem of learning mixtures of linear regressions, the goal is to learn a col-lection of signal vectors from a sequence of (possibly noisy) linear measu

rements,where each measurement is evaluated on an unknown signal drawn uniformly
fromthis collection. This setting is quite expressive and has been studied both
in termsof practical applications and for the sake of establishing theoretical
guarantees. Inthis paper, we consider the case where the signal vectors arespars
e; this generalizesthe popular compressed sensing paradigm. We improve upon the
state-of-the-artresults as follows: In the noisy case, we resolve an open questi
on of Yin et al. (IEEETransactions on Information Theory, 2019) by showing how t
o handle collectionsof more than two vectors and present the first robust recons
truction algorithm, i.e.,if the signals are not perfectly sparse, we still learn
a good sparse approximationof the signals. In the noiseless case, as well as in
the noisy case, we show how tocircumvent the need for a restrictive assumption
required in the previous work. Ourtechniques are quite different from those in t
he previous work: for the noiselesscase, we rely on a property of sparse polynom
ials and for the noisy case, we providenew connections to learning Gaussian mixt
ures and use ideas from the theory of
****
A Convex Relaxation Barrier to Tight Robustness Verification of Neural Networks
Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, Pengchuan Zhang
Verification of neural networks enables us to gauge their robustness against adv
ersarial attacks. Verification algorithms fall into two categories: exact verifi
ers that run in exponential time and relaxed verifiers that are efficient but in
complete. In this paper, we unify all existing LP-relaxed verifiers, to the best
of our knowledge, under a general convex relaxation framework. This framework w
orks for neural networks with diverse architectures and nonlinearities and cover
s both primal and dual views of neural network verification. Next, we perform la
rge-scale experiments, amounting to more than 22 CPU-years, to obtain exact solu
tion to the convex-relaxed problem that is optimal within our framework for ReLU
networks. We find the exact solution does not significantly improve upon the ga
p between PGD and existing relaxed verifiers for various networks trained normal
ly or robustly on MNIST and CIFAR datasets. Our results suggest there is an inhe
rent barrier to tight verification for the large class of methods captured by ou
r framework. We discuss possible causes of this barrier and potential future dir
ections for bypassing it.
****
A Latent Variational Framework for Stochastic Optimization
Philippe Casgrain
This paper provides a unifying theoretical framework for stochastic optimization
algorithms by means of a latent stochastic variational problem. Using technique
s from stochastic control, the solution to the variational problem is shown to b
e equivalent to that of a Forward Backward Stochastic Differential Equation (FBS
DE). By solving these equations, we recover a variety of existing adaptive stoch
astic gradient descent methods. This framework establishes a direct connection b
etween stochastic optimization algorithms and a secondary latent inference probl
em on gradients, where a prior measure on gradient observations determines the r
esulting algorithm.
****
Escaping from saddle points on Riemannian manifolds
Yue Sun, Nicolas Flammarion, Maryam Fazel
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.
****
Solving a Class of Non-Convex Min-Max Games Using Iterative First Order Methods
Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D. Lee, Meisam Razaviyayn
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.
****

The Option Keyboard: Combining Skills in Reinforcement Learning
Andre Barreto, Diana Borsa, Shaobo Hou, Gheorghe Comanici, Eser Aygün, Philippe Hamel, Daniel Toyama, Jonathan hunt, Shibl Mourad, David Silver, Doina Precup
The ability to combine known skills to create new ones may be crucial in the solution of complex reinforcement learning problems that unfold over extended periods. We argue that a robust way of combining skills is to define and manipulate them in the space of pseudo-rewards (or "cumulants"). Based on this premise, we propose a framework for combining skills using the formalism of options. We show that every deterministic option can be unambiguously represented as a cumulant defined in an extended domain. Building on this insight and on previous results on transfer learning, we show how to approximate options whose cumulants are linear combinations of the cumulants of known options. This means that, once we have learned options associated with a set of cumulants, we can instantaneously synthesise options induced by any linear combination of them, without any learning involved. We describe how this framework provides a hierarchical interface to the environment whose abstract actions correspond to combinations of basic skills. We demonstrate the practical benefits of our approach in a resource management problem and a navigation task involving a quadrupedal simulated robot.
****
On Learning Over-parameterized Neural Networks: A Functional Approximation Perspective
Lili Su, Pengkun 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.
****
Modeling Tabular data using Conditional GAN
Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, Kalyan Veeramachaneni
Modeling the probability distribution of rows in tabular data and generating realistic synthetic data is a non-trivial task. Tabular data usually contains a mix of discrete and continuous columns. Continuous columns may have multiple modes whereas discrete columns are sometimes imbalanced making the modeling difficult. Existing statistical and deep neural network models fail to properly model this type of data. We design CTGAN, which uses a conditional generative adversarial network to address these challenges. To aid in a fair and thorough comparison, we design a benchmark with 7 simulated and 8 real datasets and several Bayesian network baselines. CTGAN outperforms Bayesian methods on most of the real datasets whereas other deep learning methods could not.
****
Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates
Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, Simon Lacoste-Julien
Recent works have shown that stochastic gradient descent (SGD) achieves the fast convergence rates of full-batch gradient descent for over-parameterized models satisfying certain interpolation conditions. However, the step-size used in these works depends on unknown quantities and SGD's practical performance heavily relies on the choice of this step-size. We propose to use line-search techniques to automatically set the step-size when training models that can interpolate the data. In the interpolation setting, we prove that SGD with a stochastic variant of the classic Armijo line-search attains the deterministic convergence rates for both convex and strongly-convex functions. Under additional assumptions, SGD with Armijo line-search is shown to achieve fast convergence for non-convex functions. Furthermore, we show that stochastic extra-gradient with a Lipschitz line-search attains linear convergence for an important class of non-convex functions and saddle-point problems satisfying interpolation. To improve the proposed methods' practical performance, we give heuristics to use larger step-sizes and acceleration. We compare the proposed algorithms against numerous optimization methods on standard classification tasks using both kernel methods and deep networks. The proposed methods result in competitive performance across all models and d

atasets, while being robust to the precise choices of hyper-parameters. For mult
i-class classification using deep networks, SGD with Armijo line-search results
in both faster convergence and better generalization.
****
Tight Regret Bounds for Model-Based Reinforcement Learning with Greedy Policies
Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, Shie Mannor
State-of-the-art efficient model-based Reinforcement Learning (RL) algorithms ty
pically act by iteratively solving empirical models, i.e., by performing full-p
lanning on Markov Decision Processes (MDPs) built by the gathered experience. In
 this paper, we focus on model-based RL in the finite-state finite-horizon MDP s
etting and establish that exploring with greedy policies -- act by 1-step planni
ng -- can achieve tight minimax performance in terms of regret, O(\sqrt{HSAT}).
 Thus, full-planning in model-based RL can be avoided altogether without any per
formance degradation, and, by doing so, the computational complexity decreases b
y a factor of S. The results are based on a novel analysis of real-time dynamic
programming, then extended to model-based RL. Specifically, we generalize existi
ng algorithms that perform full-planning to such that act by 1-step planning. Fo
r these generalizations, we prove regret bounds with the same rate as their full
-planning counterparts.
****
Weighted Linear Bandits for Non-Stationary Environments
Yoan Russac, Claire Vernade, Olivier Cappé
We consider a stochastic linear bandit model in which the available actions
  correspond to arbitrary context vectors whose associated rewards
  follow a non-stationary linear regression model.
  In this setting, the unknown regression parameter is allowed to vary in time.
 To address this problem, we propose
  D-LinUCB, a novel optimistic algorithm based on discounted linear regression,
where exponential weights are used to smoothly forget
  the past.
  This involves studying the deviations of the sequential weighted least-square
s estimator under generic assumptions.
  As a by-product, we obtain novel deviation results that can be used beyond no
n-stationary environments.
   We provide theoretical guarantees on the behavior of
  D-LinUCB in both slowly-varying and abruptly-changing
  environments. We obtain an upper bound on the
  dynamic regret that is of order d BT^{1/3}T^{2/3}, where BT
  is a measure of non-stationarity (d and T being, respectively, dimension and h
orizon). This rate is known to be optimal. We
  also illustrate the empirical performance of D-LinUCB
  and compare it with recently proposed alternatives in
  simulated environments.
****
Neural Lyapunov Control
Ya-Chien Chang, Nima Roohi, Sicun Gao
We propose new methods for learning control policies and neural network Lyapunov
 functions for nonlinear control problems, with provable guarantee of stability.
 The framework consists of a learner that attempts to find the control and Lyapu
nov functions, and a falsifier that finds counterexamples to quickly guide the l
earner towards solutions. The procedure terminates when no counterexample is fou
nd by the falsifier, in which case the controlled nonlinear system is provably s
table. The approach significantly simplifies the process of Lyapunov control des
ign, provides end-to-end correctness guarantee, and can obtain much larger regio
ns of attraction than existing methods such as LQR and SOS/SDP. We show experime
nts on how the new methods obtain high-quality solutions for challenging robot c
ontrol problems such as path tracking for wheeled vehicles and humanoid robot ba
lancing.
****
Stochastic Variance Reduced Primal Dual Algorithms for Empirical Composition Opt

imization
Adithya M Devraj, Jianshu Chen
We consider a generic empirical composition optimization problem, where there ar
e empirical averages present both outside and inside nonlinear loss functions. S
uch a problem is of interest in various machine learning applications, and canno
t be directly solved by standard methods such as stochastic gradient descent (SG
D). We take a novel approach to solving this problem by reformulating the origin
al minimization objective into an equivalent min-max objective, which brings out
 all the empirical averages that are originally inside the nonlinear loss functi
ons. We exploit the rich structures of the reformulated problem and develop a st
ochastic primal-dual algorithms, SVRPDA-I, to solve the problem efficiently. We
carry out extensive theoretical analysis of the proposed algorithm, obtaining th
e convergence rate, the total computation complexity and the storage complexity.
 In particular, the algorithm is shown to converge at a linear rate when the pro
blem is strongly convex. Moreover, we also develop an approximate version of the
 algorithm, named SVRPDA-II, which further reduces the memory requirement. Final
ly, we evaluate the performance of our algorithms on several real-world benchmar
ks and experimental results show that they significantly outperform existing tec
hniques.
****
Hamiltonian Neural Networks
Samuel Greydanus, Misko Dzamba, Jason Yosinski
Even though neural networks enjoy widespread use, they still struggle to learn t
he basic laws of physics. How might we endow them with better inductive biases?
In this paper, we draw inspiration from Hamiltonian mechanics to train models th
at learn and respect exact conservation laws in an unsupervised manner. We evalu
ate our models on problems where conservation of energy is important, including
the two-body problem and pixel observations of a pendulum. Our model trains fast
er and generalizes better than a regular neural network. An interesting side eff
ect is that our model is perfectly reversible in time.
****
Better Transfer Learning with Inferred Successor Maps
Tamas Madarasz, Tim Behrens
Humans and animals show remarkable flexibility in adjusting their behaviour when
 their goals, or rewards in the environment change. While such flexibility is a
hallmark of intelligent behaviour, these multi-task scenarios remain an importan
t challenge for machine learning algorithms and neurobiological models alike. We
 investigated two approaches that could enable this flexibility: factorized repr
esentations, which abstract away general aspects of a task from those prone to c
hange, and nonparametric, memory-based approaches, which can provide a principle
d way of using similarity to past experiences to guide current behaviour. In par
ticular, we combine the successor representation (SR), that factors the value of
 actions into expected outcomes and corresponding rewards, with evaluating task
 similarity through clustering the space of rewards. The proposed algorithm inve
rts a generative model over tasks, and dynamically samples from a flexible numbe
r of distinct SR maps while accumulating evidence about the current task context
 through amortized inference. It improves SR's transfer capabilities and outperf
orms competing algorithms and baselines in settings with both known and unsignal
led rewards changes. Further, as a neurobiological model of spatial coding in th
e hippocampus, it explains important signatures of this representation, such as
the "flickering" behaviour of hippocampal maps, and trajectory-dependent place
cells (so-called splitter cells) and their dynamics. We thus provide a novel alg
orithmic approach for multi-task learning, as well as a common normative framewo
rk that links together these different characteristics of the brain's spatial re
presentation.
****
Random Quadratic Forms with Dependence: Applications to Restricted Isometry and
Beyond
Arindam Banerjee, Qilong Gu, Vidyashankar Sivakumar, Steven Z. Wu
Several important families of computational and statistical results in machine l

earning and randomized algorithms rely on uniform bounds on quadratic forms of random vectors or matrices. Such results include the Johnson-Lindenstrauss (J-L) Lemma, the Restricted Isometry Property (RIP), randomized sketching algorithms, and approximate linear algebra. The existing results critically depend on statistical independence, e.g., independent entries for random vectors, independent rows for random matrices, etc., which prevent their usage in dependent or adaptive modeling settings. In this paper, we show that such independence is in fact not needed for such results which continue to hold under fairly general dependence structures. In particular, we present uniform bounds on random quadratic forms of stochastic processes which are conditionally independent and sub-Gaussian given another (latent) process. Our setup allows general dependencies of the stochastic process on the history of the latent process and the latent process to be influenced by realizations of the stochastic process. The results are thus applicable to adaptive modeling settings and also allows for sequential design of random vectors and matrices. We also discuss stochastic process based forms of J-L, RIP, and sketching, to illustrate the generality of the results.
****

Energy-Inspired Models: Learning with Sampler-Induced Distributions

John Lawson, George Tucker, Bo Dai, Rajesh Ranganath

Energy-based models (EBMs) are powerful probabilistic models, but suffer from intractable sampling and density evaluation due to the partition function. As a result, inference in EBMs relies on approximate sampling algorithms, leading to a mismatch between the model and inference. Motivated by this, we consider the sampler-induced distribution as the model of interest and maximize the likelihood of this model. This yields a class of energy-inspired models (EIMs) that incorporate learned energy functions while still providing exact samples and tractable log-likelihood lower bounds. We describe and evaluate three instantiations of such models based on truncated rejection sampling, self-normalized importance sampling, and Hamiltonian importance sampling. These models out-perform or perform comparably to the recently proposed Learned Accept/RejectSampling algorithm and provide new insights on ranking Noise Contrastive Estimation and Contrastive Predictive Coding. Moreover, EIMs allow us to generalize a recent connection between multi-sample variational lower bounds and auxiliary variable variational inference. We show how recent variational bounds can be unified with EIMs as the variational family.
****

Data-Dependence of Plateau Phenomenon in Learning with Neural Network --- Statistical Mechanical Analysis

Yuki Yoshida, Masato Okada

The plateau phenomenon, wherein the loss value stops decreasing during the process of learning, has been reported by various researchers. The phenomenon is actively inspected in the 1990s and found to be due to the fundamental hierarchical structure of neural network models. Then the phenomenon has been thought as inevitable. However, the phenomenon seldom occurs in the context of recent deep learning. There is a gap between theory and reality. In this paper, using statistical mechanical formulation, we clarified the relationship between the plateau phenomenon and the statistical property of the data learned. It is shown that the data whose covariance has small and dispersed eigenvalues tend to make the plateau phenomenon inconspicuous.
****

Differentiable Cloth Simulation for Inverse Problems

Junbang Liang, Ming Lin, Vladlen Koltun

We propose a differentiable cloth simulator that can be embedded as a layer in deep neural networks. This approach provides an effective, robust framework for modeling cloth dynamics, self-collisions, and contacts. Due to the high dimensionality of the dynamical system in modeling cloth, traditional gradient computation for collision response can become impractical. To address this problem, we propose to compute the gradient directly using QR decomposition of a much smaller matrix. Experimental results indicate that our method can speed up backpropagation by two orders of magnitude. We demonstrate the presented approach on a number

of inverse problems, including parameter estimation and motion control for cloth.
****

Detecting Overfitting via Adversarial Examples
Roman Werpachowski, András György, Csaba Szepesvari
The repeated community-wide reuse of test sets in popular benchmark problems raises doubts about the credibility of reported test-error rates. Verifying whether a learned model is overfitted to a test set is challenging as independent test sets drawn from the same data distribution are usually unavailable, while other test sets may introduce a distribution shift. We propose a new hypothesis test that uses only the original test data to detect overfitting. It utilizes a new unbiased error estimate that is based on adversarial examples generated from the test data and importance weighting. Overfitting is detected if this error estimate is sufficiently different from the original test error rate. We develop a specialized variant of our test for multiclass image classification, and apply it to testing overfitting of recent models to the popular ImageNet benchmark. Our method correctly indicates overfitting of the trained model to the training set, but is not able to detect any overfitting to the test set, in line with other recent work on this topic.
****

Region-specific Diffeomorphic Metric Mapping
Zhengyang Shen, Francois-Xavier Vialard, Marc Niethammer
We introduce a region-specific diffeomorphic metric mapping (RDMM) registration approach. RDMM is non-parametric, estimating spatio-temporal velocity fields which parameterize the sought-for spatial transformation. Regularization of these velocity fields is necessary. In contrast to existing non-parametric registration approaches using a fixed spatially-invariant regularization, for example, the large displacement diffeomorphic metric mapping (LDDMM) model, our approach allows for spatially-varying regularization which is advected via the estimated spatio-temporal velocity field. Hence, not only can our model capture large displacements, it does so with a spatio-temporal regularizer that keeps track of how regions deform, which is a more natural mathematical formulation. We explore a family of RDMM registration approaches: 1) a registration model where regions with separate regularizations are pre-defined (e.g., in an atlas space or for distinct foreground and background regions), 2) a registration model where a general spatially-varying regularizer is estimated, and 3) a registration model where the spatially-varying regularizer is obtained via an end-to-end trained deep learning (DL) model. We provide a variational derivation of RDMM, showing that the model can assure diffeomorphic transformations in the continuum, and that LDDMM is a particular instance of RDMM. To evaluate RDMM performance we experiment 1) on synthetic 2D data and 2) on two 3D datasets: knee magnetic resonance images (MRIs) of the Osteoarthritis Initiative (OAI) and computed tomography images (CT) of the lung. Results show that our framework achieves comparable performance to state-of-the-art image registration approaches, while providing additional information via a learned spatio-temporal regularizer. Further, our deep learning approach allows for very fast RDMM and LDDMM estimations. Code is available at https://github.com/uncbiag/registration.
****

Teaching Multiple Concepts to a Forgetful Learner
Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, Adish Singla
How can we help a forgetful learner learn multiple concepts within a limited time frame? While there have been extensive studies in designing optimal schedules for teaching a single concept given a learner's memory model, existing approaches for teaching multiple concepts are typically based on heuristic scheduling techniques without theoretical guarantees. In this paper, we look at the problem from the perspective of discrete optimization and introduce a novel algorithmic framework for teaching multiple concepts with strong performance guarantees. Our framework is both generic, allowing the design of teaching schedules for different memory models, and also interactive, allowing the teacher to adapt the schedu

le to the underlying forgetting mechanisms of the learner. Furthermore, for a we
ll-known memory model, we are able to identify a regime of model parameters wher
e our framework is guaranteed to achieve high performance. We perform extensive
evaluations using simulations along with real user studies in two concrete appli
cations: (i) an educational app for online vocabulary teaching; and (ii) an app
for teaching novices how to recognize animal species from images. Our results d
emonstrate the effectiveness of our algorithm compared to popular heuristic appr
oaches.
****
Domain Generalization via Model-Agnostic Learning of Semantic Features
Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, Ben Glocker
Generalization capability to unseen domains is crucial for machine learning mode
ls when deploying to real-world conditions. We investigate the challenging probl
em of domain generalization, i.e., training a model on multi-domain source data
such that it can directly generalize to target domains with unknown statistics.
We adopt a model-agnostic learning paradigm with gradient-based meta-train and m
eta-test procedures to expose the optimization to domain shift. Further, we intr
oduce two complementary losses which explicitly regularize the semantic structur
e of the feature space. Globally, we align a derived soft confusion matrix to pr
eserve general knowledge of inter-class relationships. Locally, we promote domai
n-independent class-specific cohesion and separation of sample features with a m
etric-learning component. The effectiveness of our method is demonstrated with n
ew state-of-the-art results on two common object recognition benchmarks. Our met
hod also shows consistent improvement on a medical image segmentation task.
****
Unconstrained Monotonic Neural Networks
Antoine Wehenkel, Gilles Louppe
Monotonic neural networks have recently been proposed as a way to define inverti
ble transformations. These transformations can be combined into powerful autoreg
ressive flows that have been shown to be universal approximators of continuous p
robability distributions. Architectures that ensure monotonicity typically enfor
ce constraints on weights and activation functions, which enables invertibility
but leads to a cap on the expressiveness of the resulting transformations. In t
his work, we propose the Unconstrained Monotonic Neural Network (UMNN) architect
ure based on the insight that a function is monotonic as long as its derivative
is strictly positive. In particular, this latter condition can be enforced with
a free-form neural network whose only constraint is the positiveness of its outp
ut. We evaluate our new invertible building block within a new autoregressive f
low (UMNN-MAF) and demonstrate its effectiveness on density estimation experimen
ts. We also illustrate the ability of UMNNs to improve variational inference.
****
Efficient Identification in Linear Structural Causal Models with Instrumental Cu
tsets
Daniel Kumor, Bryant Chen, Elias Bareinboim
One of the most common mistakes made when performing data analysis is attributin
g causal meaning to regression coefficients. Formally, a causal effect can only
be computed if it is identifiable from a combination of observational data and s
tructural knowledge about the domain under investigation (Pearl, 2000, Ch. 5). B
uilding on the literature of instrumental variables (IVs), a plethora of methods
 has been developed to identify causal effects in linear systems. Almost invaria
bly, however, the most powerful such methods rely on exponential-time procedures
. In this paper, we investigate graphical conditions to allow efficient identifi
cation in arbitrary linear structural causal models (SCMs). In particular, we de
velop a method to efficiently find unconditioned instrumental subsets, which are
 generalizations of IVs that can be used to tame the complexity of many canonica
l algorithms found in the literature. Further, we prove that determining whether
 an effect can be identified with TSID (Weihs et al., 2017), a method more power
ful than unconditioned instrumental sets and other efficient identification algo
rithms, is NP-Complete. Finally, building on the idea of flow constraints, we in
troduce a new and efficient criterion called Instrumental Cutsets (IC), which is

able to solve for parameters missed by all other existing polynomial-time algorithms.
****
Temporal FiLM: Capturing Long-Range Sequence Dependencies with Feature-Wise Modulations.
Sawyer Birnbaum, Volodymyr Kuleshov, Zayd Enam, Pang Wei W. Koh, Stefano Ermon
Learning representations that accurately capture long-range dependencies in sequential inputs --- including text, audio, and genomic data --- is a key problem in deep learning. Feed-forward convolutional models capture only feature interactions within finite receptive fields while recurrent architectures can be slow and difficult to train due to vanishing gradients. Here, we propose Temporal Feature-Wise Linear Modulation (TFiLM) --- a novel architectural component inspired by adaptive batch normalization and its extensions --- that uses a recurrent neural network to alter the activations of a convolutional model. This approach expands the receptive field of convolutional sequence models with minimal computational overhead. Empirically, we find that TFiLM significantly improves the learning speed and accuracy of feed-forward neural networks on a range of generative and discriminative learning tasks, including text classification and audio super-resolution.
****
Convolution with even-sized kernels and symmetric padding
Shuang Wu, Guanrui Wang, Pei Tang, Feng Chen, Luping Shi
Compact convolutional neural networks gain efficiency mainly through depthwise convolutions, expanded channels and complex topologies, which contrarily aggravate the training process. Besides, 3x3 kernels dominate the spatial representation in these models, whereas even-sized kernels (2x2, 4x4) are rarely adopted. In this work, we quantify the shift problem occurs in even-sized kernel convolutions by an information erosion hypothesis, and eliminate it by proposing symmetric padding on four sides of the feature maps (C2sp, C4sp). Symmetric padding releases the generalization capabilities of even-sized kernels at little computational cost, making them outperform 3x3 kernels in image classification and generation tasks. Moreover, C2sp obtains comparable accuracy to emerging compact models with much less memory and time consumption during training.
Symmetric padding coupled with even-sized convolutions can be neatly implemented into existing frameworks, providing effective elements for architecture designs, especially on online and continual learning occasions where training efforts are emphasized.
****
Inducing brain-relevant bias in natural language processing models
Dan Schwartz, Mariya Toneva, Leila Wehbe
Progress in natural language processing (NLP) models that estimate representations of word sequences has recently been leveraged to improve the understanding of language processing in the brain. However, these models have not been specifically designed to capture the way the brain represents language meaning. We hypothesize that fine-tuning these models to predict recordings of brain activity of people reading text will lead to representations that encode more brain-activity-relevant language information. We demonstrate that a version of BERT, a recently introduced and powerful language model, can improve the prediction of brain activity after fine-tuning. We show that the relationship between language and brain activity learned by BERT during this fine-tuning transfers across multiple participants. We also show that, for some participants, the fine-tuned representations learned from both magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) are better for predicting fMRI than the representations learned from fMRI alone, indicating that the learned representations capture brain-activity-relevant information that is not simply an artifact of the modality. While changes to language representations help the model predict brain activity, they also do not harm the model's ability to perform downstream NLP tasks. Our findings are notable for research on language understanding in the brain.
****
SMILe: Scalable Meta Inverse Reinforcement Learning through Context-Conditional

Policies
Seyed Kamyar Seyed Ghasemipour, Shixiang (Shane) Gu, Richard Zemel
Imitation Learning (IL) has been successfully applied to complex sequential decision-making problems where standard Reinforcement Learning (RL) algorithms fail. A number of recent methods extend IL to few-shot learning scenarios, where a meta-trained policy learns to quickly master new tasks using limited demonstrations. However, although Inverse Reinforcement Learning (IRL) often outperforms Behavioral Cloning (BC) in terms of imitation quality, most of these approaches build on BC due to its simple optimization objective. In this work, we propose SMILe, a scalable framework for Meta Inverse Reinforcement Learning (Meta-IRL) based on maximum entropy IRL, which can learn high-quality policies from few demonstrations. We examine the efficacy of our method on a variety of high-dimensional simulated continuous control tasks and observe that SMILe significantly outperforms Meta-BC. Furthermore, we observe that SMILe performs comparably or outperforms Meta-DAgger, while being applicable in the state-only setting and not requiring online experts. To our knowledge, our approach is the first efficient method for Meta-IRL that scales to the function approximator setting. For datasets and reproducing results please refer to https://github.com/KamyarGh/rlswiss/blob/master/reproducing/smilepaper.md .
****
Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model
Erik Nijkamp, Mitch Hill, Song-Chun Zhu, Ying Nian Wu
This paper studies a curious phenomenon in learning energy-based model (EBM) using MCMC. In each learning iteration, we generate synthesized examples by running a non-convergent, non-mixing, and non-persistent short-run MCMC toward the current model, always starting from the same initial distribution such as uniform noise distribution, and always running a fixed number of MCMC steps. After generating synthesized examples, we then update the model parameters according to the maximum likelihood learning gradient, as if the synthesized examples are fair samples from the current model. We treat this non-convergent short-run MCMC as a learned generator model or a flow model. We provide arguments for treating the learned non-convergent short-run MCMC as a valid model. We show that the learned short-run MCMC is capable of generating realistic images. More interestingly, unlike traditional EBM or MCMC, the learned short-run MCMC is capable of reconstructing observed images and interpolating between images, like generator or flow models. The code can be found in the Appendix.
****
Exploring Unexplored Tensor Network Decompositions for Convolutional Neural Networks
Kohei Hayashi, Taiki Yamaguchi, Yohei Sugawara, Shin-ichi Maeda
Tensor decomposition methods are widely used for model compression and fast inference in convolutional neural networks (CNNs). Although many decompositions are conceivable, only CP decomposition and a few others have been applied in practice, and no extensive comparisons have been made between available methods. Previous studies have not determined how many decompositions are available, nor which of them is optimal. In this study, we first characterize a decomposition class specific to CNNs by adopting a flexible graphical notation. The class includes such well-known CNN modules as depthwise separable convolution layers and bottleneck layers, but also previously unknown modules with nonlinear activations. We also experimentally compare the tradeoff between prediction accuracy and time/space complexity for modules found by enumerating all possible decompositions, or by using a neural architecture search. We find some nonlinear decompositions outperform existing ones.
****
Interval timing in deep reinforcement learning agents
Ben Deverett, Ryan Faulkner, Meire Fortunato, Gregory Wayne, Joel Z. Leibo
The measurement of time is central to intelligent behavior. We know that both animals and artificial agents can successfully use temporal dependencies to select actions. In artificial agents, little work has directly addressed (1) which architectural components are necessary for successful development of this ability,

(2) how this timing ability comes to be represented in the units and actions of the agent, and (3) whether the resulting behavior of the system converges on solutions similar to those of biology. Here we studied interval timing abilities in deep reinforcement learning agents trained end-to-end on an interval reproduction paradigm inspired by experimental literature on mechanisms of timing. We characterize the strategies developed by recurrent and feedforward agents, which both succeed at temporal reproduction using distinct mechanisms, some of which bear specific and intriguing similarities to biological systems. These findings advance our understanding of how agents come to represent time, and they highlight the value of experimentally inspired approaches to characterizing agent abilities.
****
Shaping Belief States with Generative Environment Models for RL
Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, Aaron van den Oord
When agents interact with a complex environment, they must form and maintain beliefs about the relevant aspects of that environment. We propose a way to efficiently train expressive generative models in complex environments. We show that a predictive algorithm with an expressive generative model can form stable belief-states in visually rich and dynamic 3D environments. More precisely, we show that the learned representation captures the layout of the environment as well as the position and orientation of the agent. Our experiments show that the model substantially improves data-efficiency on a number of reinforcement learning (RL) tasks compared with strong model-free baseline agents. We find that predicting multiple steps into the future (overshooting), in combination with an expressive generative model, is critical for stable representations to emerge. In practice, using expressive generative models in RL is computationally expensive and we propose a scheme to reduce this computational burden, allowing us to build agents that are competitive with model-free baselines.
****
Uncertainty-based Continual Learning with Adaptive Regularization
Hongjoon Ahn, Sungmin Cha, Donggyu Lee, Taesup Moon
We introduce a new neural network-based continual learning algorithm, dubbed as Uncertainty-regularized Continual Learning (UCL), which builds on traditional Bayesian online learning framework with variational inference. We focus on two significant drawbacks of the recently proposed regularization-based methods: a) considerable additional memory cost for determining the per-weight regularization strengths and b) the absence of gracefully forgetting scheme, which can prevent performance degradation in learning new tasks. In this paper, we show UCL can solve these two problems by introducing a fresh interpretation on the Kullback-Leibler (KL) divergence term of the variational lower bound for Gaussian mean-field approximation. Based on the interpretation, we propose the notion of node-wise uncertainty, which drastically reduces the number of additional parameters for implementing per-weight regularization. Moreover, we devise two additional regularization terms that enforce \emph{stability} by freezing important parameters for past tasks and allow \emph{plasticity} by controlling the actively learning parameters for a new task. Through extensive experiments, we show UCL convincingly outperforms most of recent state-of-the-art baselines not only on popular supervised learning benchmarks, but also on challenging lifelong reinforcement learning tasks. The source code of our algorithm is available at https://github.com/csm9493/UCL.
****
Implicit Posterior Variational Inference for Deep Gaussian Processes
Haibin YU, Yizhou Chen, Bryan Kian Hsiang Low, Patrick Jaillet, Zhongxiang Dai
A multi-layer deep Gaussian process (DGP) model is a hierarchical composition of GP models with a greater expressive power. Exact DGP inference is intractable, which has motivated the recent development of deterministic and stochastic approximation methods. Unfortunately, the deterministic approximation methods yield a biased posterior belief while the stochastic one is computationally costly. This paper presents an implicit posterior variational inference (IPVI) framework fo

r DGPs that can ideally recover an unbiased posterior belief and still preserve time efficiency. Inspired by generative adversarial networks, our IPVI framework achieves this by casting the DGP inference problem as a two-player game in which a Nash equilibrium, interestingly, coincides with an unbiased posterior belief. This consequently inspires us to devise a best-response dynamics algorithm to search for a Nash equilibrium (i.e., an unbiased posterior belief). Empirical evaluation shows that IPVI outperforms the state-of-the-art approximation methods for DGPs.
****
Are Sixteen Heads Really Better than One?
Paul Michel, Omer Levy, Graham Neubig
Multi-headed attention is a driving force behind recent state-of-the-art NLP models.
By applying multiple attention mechanisms in parallel, it can express sophisticated functions beyond the simple weighted average.
However we observe that, in practice, a large proportion of attention heads can be removed at test time without significantly impacting performance, and that some layers can even be reduced to a single head.
Further analysis on machine translation models reveals that the self-attention layers can be significantly pruned, while the encoder-decoder layers are more dependent on multi-headedness.
****
Model Compression with Adversarial Robustness: A Unified Optimization Framework
Shupeng Gui, Haotao Wang, Haichuan Yang, Chen Yu, Zhangyang Wang, Ji Liu
Deep model compression has been extensively studied, and state-of-the-art methods can now achieve high compression ratios with minimal accuracy loss. This paper studies model compression through a different lens: could we compress models without hurting their robustness to adversarial attacks, in addition to maintaining accuracy? Previous literature suggested that the goals of robustness and compactness might sometimes contradict. We propose a novel Adversarially Trained Model Compression (ATMC) framework. ATMC constructs a unified constrained optimization formulation, where existing compression means (pruning, factorization, quantization) are all integrated into the constraints. An efficient algorithm is then developed. An extensive group of experiments are presented, demonstrating that ATMC obtains remarkably more favorable trade-off among model size, accuracy and robustness, over currently available alternatives in various settings. The codes are publicly available at: https://github.com/shupenggui/ATMC.
****
Subspace Attack: Exploiting Promising Subspaces for Query-Efficient Black-box Attacks
Yiwen Guo, Ziang Yan, Changshui Zhang
Unlike the white-box counterparts that are widely studied and readily accessible, adversarial examples in black-box settings are generally more Herculean on account of the difficulty of estimating gradients. Many methods achieve the task by issuing numerous queries to target classification systems, which makes the whole procedure costly and suspicious to the systems. In this paper, we aim at reducing the query complexity of black-box attacks in this category. We propose to exploit gradients of a few reference models which arguably span some promising search subspaces. Experimental results show that, in comparison with the state-of-the-arts, our method can gain up to 2x and 4x reductions in the requisite mean and medium numbers of queries with much lower failure rates even if the reference models are trained on a small and inadequate dataset disjoint to the one for training the victim model. Code and models for reproducing our results will be made publicly available.
****
Combinatorial Bayesian Optimization using the Graph Cartesian Product
Changyong Oh, Jakub Tomczak, Efstratios Gavves, Max Welling
This paper focuses on Bayesian Optimization (BO) for objectives on combinatorial search spaces, including ordinal and categorical variables. Despite the abundance

of potential applications of Combinatorial BO, including chipset configuration search and neural architecture search, only a handful of methods have been proposed. We introduce COMBO, a new Gaussian Process (GP) BO. COMBO quantifies “smoothness” of functions on combinatorial search spaces by utilizing a combinatorial graph. The vertex set of the combinatorial graph consists of all possible joint assignments of the variables, while edges are constructed using the graph Cartesian product of the sub-graphs that represent the individual variables. On this combinatorial graph, we propose an ARD diffusion kernel with which the GP is able to model high-order interactions between variables leading to better performance. Moreover, using the Horseshoe prior for the scale parameter in the ARD diffusion kernel results in an effective variable selection procedure, making COMBO suitable for high dimensional problems. Computationally, in COMBO the graph Cartesian product allows the Graph Fourier Transform calculation to scale linearly instead of exponentially.We validate COMBO in a wide array of realistic benchmarks, including weighted maximum satisfiability problems and neural architecture search. COMBO outperforms consistently the latest state-of-the-art while maintaining computational and statistical efficiency
****
Sample Adaptive MCMC
Michael Zhu
For MCMC methods like Metropolis-Hastings, tuning the proposal distribution is important in practice for effective sampling from the target distribution \pi. In this paper, we present Sample Adaptive MCMC (SA-MCMC), a MCMC method based on a reversible Markov chain for \pi^{\otimes N} that uses an adaptive proposal distribution based on the current state of N points and a sequential substitution procedure with one new likelihood evaluation per iteration and at most one updated point each iteration. The SA-MCMC proposal distribution automatically adapts within its parametric family to best approximate the target distribution, so in contrast to many existing MCMC methods, SA-MCMC does not require any tuning of the proposal distribution. Instead, SA-MCMC only requires specifying the initial state of N points, which can often be chosen a priori, thereby automating the entire sampling procedure with no tuning required. Experimental results demonstrate the fast adaptation and effective sampling of SA-MCMC.
****
Tree-Sliced Variants of Wasserstein Distances
Tam Le, Makoto Yamada, Kenji Fukumizu, Marco Cuturi
Optimal transport (\OT) theory defines a powerful set of tools to compare probability distributions. \OT~suffers however from a few drawbacks, computational and statistical, which have encouraged the proposal of several regularized variants of OT in the recent literature, one of the most notable being the \textit{sliced} formulation, which exploits the closed-form formula between univariate distributions by projecting high-dimensional measures onto random lines. We consider in this work a more general family of ground metrics, namely \textit{tree metrics}, which also yield fast closed-form computations and negative definite, and of which the sliced-Wasserstein distance is a particular case (the tree is a chain). We propose the tree-sliced Wasserstein distance, computed by averaging the Wasserstein distance between these measures using random tree metrics, built adaptively in either low or high-dimensional spaces. Exploiting the negative definiteness of that distance, we also propose a positive definite kernel, and test it against other baselines on a few benchmark tasks.
****
Integrating Markov processes with structural causal modeling enables counterfactual inference in complex systems
Robert Ness, Kaushal Paneri, Olga Vitek
This manuscript contributes a general and practical framework for casting a Markov process model of a system at equilibrium as a structural causal model, and ca

rrying out counterfactual inference. Markov processes mathematically describe the mechanisms in the system, and predict the system's equilibrium behavior upon intervention, but do not support counterfactual inference. In contrast, structural causal models support counterfactual inference, but do not identify the mechanisms. This manuscript leverages the benefits of both approaches. We define the structural causal models in terms of the parameters and the equilibrium dynamics of the Markov process models, and counterfactual inference flows from these settings. The proposed approach alleviates the identifiability drawback of the structural causal models, in that the counterfactual inference is consistent with the counterfactual trajectories simulated from the Markov process model. We showcase the benefits of this framework in case studies of complex biomolecular systems with nonlinear dynamics. We illustrate that, in presence of Markov process model misspecification, counterfactual inference leverages prior data, and therefore estimates the outcome of an intervention more accurately than a direct simulation.
****
An Adaptive Empirical Bayesian Method for Sparse Deep Learning
Wei Deng, Xiao Zhang, Faming Liang, Guang Lin
We propose a novel adaptive empirical Bayesian (AEB) method for sparse deep learning, where the sparsity is ensured via a class of self-adaptive spike-and-slab priors. The proposed method works by alternatively sampling from an adaptive hierarchical posterior distribution using stochastic gradient Markov Chain Monte Carlo (MCMC) and smoothly optimizing the hyperparameters using stochastic approximation (SA). The convergence of the proposed method to the asymptotically correct distribution is established under mild conditions. Empirical applications of the proposed method lead to the state-of-the-art performance on MNIST and Fashion MNIST with shallow convolutional neural networks (CNN) and the state-of-the-art compression performance on CIFAR10 with Residual Networks. The proposed method also improves resistance to adversarial attacks.
****
Topology-Preserving Deep Image Segmentation
Xiaoling Hu, Fuxin Li, Dimitris Samaras, Chao Chen
Segmentation algorithms are prone to make topological errors on fine-scale struc-
tures, e.g., broken connections. We propose a novel method that learns to segment with correct topology. In particular, we design a continuous-valued loss function that enforces a segmentation to have the same topology as the ground truth, i.e.,having the same Betti number. The proposed topology-preserving loss function is differentiable and can be incorporated into end-to-end training of a deep neural network. Our method achieves much better performance on the Betti number error, which directly accounts for the topological correctness. It also performs superior on other topology-relevant metrics, e.g., the Adjusted Rand Index and the Variation of Information, without sacrificing per-pixel accuracy. We illustrate the effectiveness of the proposed method on a broad spectrum of natural and biomedical datasets.
****
Stacked Capsule Autoencoders
Adam Kosiorek, Sara Sabour, Yee Whye Teh, Geoffrey E. Hinton
Objects are composed of a set of geometrically organized parts. We introduce an unsupervised capsule autoencoder (SCAE), which explicitly uses geometric relationships between parts to reason about objects.
Since these relationships do not depend on the viewpoint, our model is robust to viewpoint changes.
SCAE consists of two stages.
In the first stage, the model predicts presences and poses of part templates directly from the image and tries to reconstruct the image by appropriately arranging the templates.
In the second stage, the SCAE predicts parameters of a few object capsules, which are then used to reconstruct part poses.
Inference in this model is amortized and performed by off-the-shelf neural encod

ers, unlike in previous capsule networks.
We find that object capsule presences are highly informative of the object class, which leads to state-of-the-art results for unsupervised classification on SVHN (55%) and MNIST (98.7%).
****
Progressive Augmentation of GANs
Dan Zhang, Anna Khoreva
Training of Generative Adversarial Networks (GANs) is notoriously fragile, requiring to maintain a careful balance between the generator and the discriminator in order to perform well. To mitigate this issue we introduce a new regularization technique - progressive augmentation of GANs (PA-GAN). The key idea is to gradually increase the task difficulty of the discriminator by progressively augmenting its input or feature space, thus enabling continuous learning of the generator. We show that the proposed progressive augmentation preserves the original GAN objective, does not compromise the discriminator's optimality and encourages a healthy competition between the generator and discriminator, leading to the better-performing generator. We experimentally demonstrate the effectiveness of PA-GAN across different architectures and on multiple benchmarks for the image synthesis task, on average achieving 3 point improvement of the FID score.
****
Online sampling from log-concave distributions
Holden Lee, 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.
****
Practical Two-Step Lookahead Bayesian Optimization
Jian Wu, Peter Frazier
Expected improvement and other acquisition functions widely used in Bayesian optimization use a "one-step" assumption: they value objective function evaluations assuming no future evaluations will be performed. Because we usually evaluate over multiple steps, this assumption may leave substantial room for improvement. Existing theory gives acquisition functions looking multiple steps in the future but calculating them requires solving a high-dimensional continuous-state continuous-action Markov decision process (MDP). Fast exact solutions of this MDP remain out of reach of today's methods. As a result, previous two- and multi-step lookahead Bayesian optimization algorithms are either too expensive to implement in most practical settings or resort to heuristics that may fail to fully realize the promise of two-step lookahead. This paper proposes a computationally efficient algorithm that provides an accurate solution to the two-step lookahead Bayesian optimization problem in seconds to at most several minutes of computation per batch of evaluations. The resulting acquisition function provides increased query efficiency and robustness compared with previous two- and multi-step lookahead methods in both single-threaded and batch experiments. This unlocks the value of two-step lookahead in practice. We demonstrate the value of our algorithm with extensive experiments on synthetic test functions and real-world problems.
****
Generalized Block-Diagonal Structure Pursuit: Learning Soft Latent Task Assignment against Negative Transfer
Zhiyong Yang, Qianqian Xu, Yangbangyan Jiang, Xiaochun Cao, Qingming 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.
****
Regret Bounds for Thompson Sampling in Episodic Restless Bandit Problems
Young Hun Jung, 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.
****
Adaptive Sequence Submodularity
Marko Mitrovic, Ehsan Kazemi, Moran Feldman, Andreas Krause, Amin Karbasi
In many machine learning applications, one needs to interactively select a sequence of items (e.g., recommending movies based on a user's feedback) or make sequential decisions in a certain order (e.g., guiding an agent through a series of states). Not only do sequences already pose a dauntingly large search space, but we must also take into account past observations, as well as the uncertainty of future outcomes. Without further structure, finding an optimal sequence is notoriously challenging, if not completely intractable. In this paper, we view the problem of adaptive and sequential decision making through the lens of submodularity and propose an adaptive greedy policy with strong theoretical guarantees. Additionally, to demonstrate the practical utility of our results, we run experiments on Amazon product recommendation and Wikipedia link prediction tasks.
****
N-Gram Graph: Simple Unsupervised Representation for Graphs, with Applications to Molecules
Shengchao Liu, Mehmet F. Demirel, Yingyu Liang
Machine learning techniques have recently been adopted in various applications in medicine, biology, chemistry, and material engineering. An important task is to predict the properties of molecules, which serves as the main subroutine in many downstream applications such as virtual screening and drug design. Despite the increasing interest, the key challenge is to construct proper representations of molecules for learning algorithms. This paper introduces the N-gram graph, a simple unsupervised representation for molecules. The method first embeds the vertices in the molecule graph. It then constructs a compact representation for the graph by assembling the vertex embeddings in short walks in the graph, which we show is equivalent to a simple graph neural network that needs no training. The representations can thus be efficiently computed and then used with supervised learning methods for prediction. Experiments on 60 tasks from 10 benchmark datasets demonstrate its advantages over both popular graph neural networks and traditional representation methods. This is complemented by theoretical analysis showing its strong representation and prediction power.
****
The spiked matrix model with generative priors
Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, Lenka Zdeborová
Using a low-dimensional parametrization of signals is a generic and powerful way to enhance performance in signal processing and statistical inference. A very popular and widely explored type of dimensionality reduction is sparsity; another type is generative modelling of signal distributions. Generative models based on neural networks, such as GANs or variational auto-encoders, are particularly performant and are gaining on applicability. In this paper we study spiked matrix models, where a low-rank matrix is observed through a noisy channel. This problem with sparse structure of the spikes has attracted broad attention in the past literature. Here, we replace the sparsity assumption by generative modelling, and investigate the consequences on statistical and algorithmic properties. We analyze the Bayes-optimal performance under specific generative models for the spike. In contrast with the sparsity assumption, we do not observe regions of parameters where statistical performance is superior to the best known algorithmic performance. We show that in the analyzed cases the approximate message passing algorithm is able to reach optimal performance. We also design enhanced spectral algorithms and analyze their performance and thresholds using random matrix theory, showing their superiority to the classical principal component analysis. We complement our theoretical results by illustrating the performance of the spectral algorithms when the spikes come from real datasets.
****
The Step Decay Schedule: A Near Optimal, Geometrically Decaying Learning Rate Pr

ocedure For Least Squares
Rong Ge, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli
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 and Improving Layer Normalization
Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, Junyang Lin
Layer normalization (LayerNorm) is a technique to normalize the distributions of intermediate layers. It enables smoother gradients, faster training, and better generalization accuracy. However, it is still unclear where the effectiveness stems from. In this paper, our main contribution is to take a step further in understanding LayerNorm. Many of previous studies believe that the success of LayerNorm comes from forward normalization. Unlike them, we find that the derivatives of the mean and variance are more important than forward normalization by re-centering and re-scaling backward gradients. Furthermore, we find that the parameters of LayerNorm, including the bias and gain, increase the risk of over-fitting and do not work in most cases. Experiments show that a simple version of LayerNorm (LayerNorm-simple) without the bias and gain outperforms LayerNorm on four datasets. It obtains the state-of-the-art performance on En-Vi machine translation.
To address the over-fitting problem, we propose a new normalization method, Adaptive Normalization (AdaNorm), by replacing the bias and gain with a new transformation function. Experiments show that AdaNorm demonstrates better results than LayerNorm on seven out of eight datasets.
****
Generative Modeling by Estimating Gradients of the Data Distribution
Yang Song, Stefano Ermon
We introduce a new generative model where samples are produced via Langevin dynamics using gradients of the data distribution estimated with score matching. Because gradients can be ill-defined and hard to estimate when the data resides on low-dimensional manifolds, we perturb the data with different levels of Gaussian noise, and jointly estimate the corresponding scores, i.e., the vector fields of gradients of the perturbed data distribution for all noise levels. For sampling, we propose an annealed Langevin dynamics where we use gradients corresponding to gradually decreasing noise levels as the sampling process gets closer to the data manifold. Our framework allows flexible model architectures, requires no sampling during training or the use of adversarial methods, and provides a learning objective that can be used for principled model comparisons. Our models produce samples
comparable to GANs on MNIST, CelebA and CIFAR-10 datasets, achieving a new state-of-the-art inception score of 8.87 on CIFAR-10. Additionally, we demonstrate that our models learn effective representations via image inpainting experiments.
****
Hypothesis Set Stability and Generalization
Dylan J. Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, Karthik Sridharan
We present a study of generalization for data-dependent hypothesis sets. We give a general learning guarantee for data-dependent hypothesis sets based on a notion of transductive Rademacher complexity. Our main result is a generalization bound for data-dependent hypothesis sets expressed in terms of a notion of hypothesis set stability and a notion of Rademacher complexity for data-dependent hypothesis sets that we introduce. This bound admits as special cases both standard Rademacher complexity bounds and algorithm-dependent uniform stability bounds. We also illustrate the use of these learning bounds in the analysis of several scenarios.
****

Balancing Efficiency and Fairness in On-Demand Ridesourcing
Nixie S. Lesmana, Xuan Zhang, Xiaohui Bei
We investigate the problem of assigning trip requests to available vehicles in on-demand ridesourcing. Much of the literature has focused on maximizing the total value of served requests, achieving efficiency on the passengers' side. However, such solutions may result in some drivers being assigned to insufficient or undesired trips, therefore losing fairness from the drivers' perspective.
****
Backprop with Approximate Activations for Memory-efficient Network Training
Ayan Chakrabarti, Benjamin Moseley
Training convolutional neural network models is memory intensive since back-propagation requires storing activations of all intermediate layers. This presents a practical concern when seeking to deploy very deep architectures in production, especially when models need to be frequently re-trained on updated datasets. In this paper, we propose a new implementation for back-propagation that significantly reduces memory usage, by enabling the use of approximations with negligible computational cost and minimal effect on training performance. The algorithm reuses common buffers to temporarily store full activations and compute the forward pass exactly. It also stores approximate per-layer copies of activations, at significant memory savings, that are used in the backward pass. Compared to simply approximating activations within standard back-propagation, our method limits accumulation of errors across layers. This allows the use of much lower-precision approximations without affecting training accuracy. Experiments on CIFAR-10, CIFAR-100, and ImageNet show that our method yields performance close to exact training, while storing activations compactly with as low as 4-bit precision.
****
Learning to Screen
Alon Cohen, Avinatan Hassidim, Haim Kaplan, Yishay Mansour, Shay Moran
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 coupled autoencoder approach for multi-modal analysis of cell types
Rohan Gala, Nathan Gouwens, Zizhen Yao, Agata Budzillo, Osnat Penn, Bosiljka Tasic, Gabe Murphy, Hongkui Zeng, Uygar Sümbül
Recent developments in high throughput profiling of individual neurons have spurred data driven exploration of the idea that there exist natural groupings of neurons referred to as cell types. The promise of this idea is that the immense complexity of brain circuits can be reduced, and effectively studied by means of interactions between cell types. While clustering of neuron populations based on a particular data modality can be used to define cell types, such definitions are often inconsistent across different characterization modalities. We pose this issue of cross-modal alignment as an optimization problem and develop an approach based on coupled training of autoencoders as a framework for such analyses. We apply this framework to a Patch-seq dataset consisting of transcriptomic and electrophysiological profiles for the same set of neurons to study consistency of representations across modalities, and evaluate cross-modal data prediction ability. We explore the problem where only a subset of neurons is characterized with more than one modality, and demonstrate that representations learned by coupled autoencoders can be used to identify types sampled only by a single modality.
****
Meta-Inverse Reinforcement Learning with Probabilistic Context Variables
Lantao Yu, Tianhe Yu, Chelsea Finn, Stefano Ermon
Reinforcement learning demands a reward function, which is often difficult to provide or design in real world applications. While inverse reinforcement learning (IRL) holds promise for automatically learning reward functions from demonstrations, several major challenges remain. First, existing IRL methods learn reward functions from scratch, requiring large numbers of demonstrations to correctly infer the reward for each task the agent may need to perform. Second, and more su

btly, existing methods typically assume demonstrations for one, isolated behavio r or task, while in practice, it is significantly more natural and scalable to p rovide datasets of heterogeneous behaviors. To this end, we propose a deep laten t variable model that is capable of learning rewards from unstructured, multi-ta sk demonstration data, and critically, use this experience to infer robust rewar ds for new, structurally-similar tasks from a single demonstration. Our experime nts on multiple continuous control tasks demonstrate the effectiveness of our ap proach compared to state-of-the-art imitation and inverse reinforcement learning methods.
****

Precision-Recall Balanced Topic Modelling
Seppo Virtanen, Mark Girolami
Topic models are becoming increasingly relevant probabilistic models for dimensi onality reduction of text data, inferring topics that capture meaningful themes of frequently co-occurring terms. We formulate topic modelling as an information retrieval task, where the goal is, based on the latent topic representation, to capture relevant term co-occurrence patterns. We evaluate performance for this task rigorously with regard to two types of errors, false negatives and positive s, based on the well-known precision-recall trade-off and provide a statistical model that allows the user to balance between the contributions of the different error types. When the user focuses solely on the contribution of false negative s ignoring false positives altogether our proposed model reduces to a standard t opic model. Extensive experiments demonstrate the proposed approach is effective and infers more coherent topics than existing related approaches.
****

Exact inference in structured prediction
Kevin Bello, Jean Honorio
Structured prediction can be thought of as a simultaneous prediction of multiple labels.
This is often done by maximizing a score function on the space of labels, which decomposes as a sum of pairwise and unary potentials.
The above is naturally modeled with a graph, where edges and vertices are relate d to pairwise and unary potentials, respectively.
We consider the generative process proposed by Globerson et al. (2015) and apply it to general connected graphs.
We analyze the structural conditions of the graph that allow for the exact recov ery of the labels.
Our results show that exact recovery is possible and achievable in polynomial ti me for a large class of graphs.
In particular, we show that graphs that are bad expanders can be exactly recover ed by adding small edge perturbations coming from the \Erdos-\Renyi model.
Finally, as a byproduct of our analysis, we provide an extension of Cheeger's in equality.
****

Practical and Consistent Estimation of f-Divergences
Paul Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, Ilya O. Tols tikhin
The estimation of an f-divergence between two probability distributions based on samples is a fundamental problem in statistics and machine learning. Most works study this problem under very weak assumptions, in which case it is provably har d.
We consider the case of stronger structural assumptions that are commonly satisf ied
in modern machine learning, including representation learning and generative
modelling with autoencoder architectures. Under these assumptions we propose and study an estimator that can be easily implemented, works well in high dimensions ,
and enjoys faster rates of convergence. We verify the behavior of our estimator empirically in both synthetic and real-data experiments, and discuss its direct implications for total correlation, entropy, and mutual information estimation.

****
Policy Poisoning in Batch Reinforcement Learning and Control
Yuzhe Ma, Xuezhou Zhang, Wen Sun, Jerry Zhu
We study a security threat to batch reinforcement learning and control where the attacker aims to poison the learned policy. The victim is a reinforcement learner / controller which first estimates the dynamics and the rewards from a batch data set, and then solves for the optimal policy with respect to the estimates. The attacker can modify the data set slightly before learning happens, and wants to force the learner into learning a target policy chosen by the attacker. We present a unified framework for solving batch policy poisoning attacks, and instantiate the attack on two standard victims: tabular certainty equivalence learner in reinforcement learning and linear quadratic regulator in control. We show that both instantiation result in a convex optimization problem on which global optimality is guaranteed, and provide analysis on attack feasibility and attack cost. Experiments show the effectiveness of policy poisoning attacks.
****
R2D2: Reliable and Repeatable Detector and Descriptor
Jerome Revaud, Cesar De Souza, Martin Humenberger, Philippe Weinzaepfel
Interest point detection and local feature description are fundamental steps in many computer vision applications. Classical approaches are based on a detect-then-describe paradigm where separate handcrafted methods are used to first identify repeatable keypoints and then represent them with a local descriptor. Neural networks trained with metric learning losses have recently caught up with these techniques, focusing on learning repeatable saliency maps for keypoint detection or learning descriptors at the detected keypoint locations. In this work, we argue that repeatable regions are not necessarily discriminative and can therefore lead to select suboptimal keypoints. Furthermore, we claim that descriptors should be learned only in regions for which matching can be performed with high confidence.
We thus propose to jointly learn keypoint detection and description together with a predictor of the local descriptor discriminativeness. This allows to avoid ambiguous areas, thus leading to reliable keypoint detection and description. Our detection-and-description approach simultaneously outputs sparse, repeatable and reliable keypoints that outperforms state-of-the-art detectors and descriptors on the HPatches dataset and on the recent Aachen Day-Night localization benchmark.
****
First Order Motion Model for Image Animation
Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, Nicu Sebe
Image animation consists of generating a video sequence so that an object in a source image is animated according to the motion of a driving video. Our framework addresses this problem without using any annotation or prior information about the specific object to animate. Once trained on a set of videos depicting objects of the same category (e.g. faces, human bodies), our method can be applied to any object of this class. To achieve this, we decouple appearance and motion information using a self-supervised formulation. To support complex motions, we use a representation consisting of a set of learned keypoints along with their local affine transformations. A generator network models occlusions arising during target motions and combines the appearance extracted from the source image and the motion derived from the driving video. Our framework scores best on diverse benchmarks and on a variety of object categories.
****
Scalable inference of topic evolution via models for latent geometric structures
Mikhail Yurochkin, Zhiwei Fan, Aritra Guha, Paraschos Koutris, XuanLong Nguyen
We develop new models and algorithms for learning the temporal dynamics of the topic polytopes and related geometric objects that arise in topic model based inference. Our model is nonparametric Bayesian and the corresponding inference algorithm is able to discover new topics as the time progresses. By exploiting the connection between the modeling of topic polytope evolution, Beta-Bernoulli proce

ss and the Hungarian matching algorithm, our method is shown to be several order
s of magnitude faster than existing topic modeling approaches, as demonstrated b
y experiments working with several million documents in under two dozens of minu
tes.
****
Anti-efficient encoding in emergent communication
Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, Marco Baroni
Despite renewed interest in emergent language simulations with
  neural networks, little is known about the basic properties of the
  induced code, and how they compare to human language. One
  fundamental characteristic of the latter, known as Zipf's Law of
  Abbreviation (ZLA), is that more frequent words are efficiently
  associated to shorter strings. We study whether the same pattern
  emerges when two neural networks, a speaker'' and alistener'',
  are trained to play a signaling game. Surprisingly, we find that
  networks develop an \emph{anti-efficient} encoding scheme,
  in which the most frequent inputs are associated to the longest messages,
  and messages in general are skewed towards the maximum length threshold.
  This anti-efficient code appears easier to discriminate for the listener,
  and, unlike in human communication, the speaker does not impose a
  contrasting least-effort pressure towards brevity. Indeed, when the
  cost function includes a penalty for longer messages, the resulting
  message distribution starts respecting ZLA. Our analysis stresses
  the importance of studying the basic features of emergent
  communication in a highly controlled setup, to ensure the latter
  will not strand too far from human language. Moreover, we present a
  concrete illustration of how different functional pressures can lead
  to successful communication codes that lack basic properties of
  human language, thus highlighting the role such pressures play in
  the latter.
****
Improving Black-box Adversarial Attacks with a Transfer-based Prior
Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu
We consider the black-box adversarial setting, where the adversary has to genera
te adversarial perturbations without access to the target models to compute grad
ients. Previous methods tried to approximate the gradient either by using a tran
sfer gradient of a surrogate white-box model, or based on the query feedback. Ho
wever, these methods often suffer from low attack success rates or poor query ef
ficiency since it is non-trivial to estimate the gradient in a high-dimensional
space with limited information. To address these problems, we propose a prior-gu
ided random gradient-free (P-RGF) method to improve black-box adversarial attack
s, which takes the advantage of a transfer-based prior and the query information
 simultaneously. The transfer-based prior given by the gradient of a surrogate m
odel is appropriately integrated into our algorithm by an optimal coefficient de
rived by a theoretical analysis. Extensive experiments demonstrate that our meth
od requires much fewer queries to attack black-box models with higher success ra
tes compared with the alternative state-of-the-art methods.
****
REM: From Structural Entropy to Community Structure Deception
Yiwei Liu, Jiamou Liu, Zijian Zhang, Liehuang Zhu, Angsheng Li
This paper focuses on the privacy risks of disclosing the community structure in
 an online social network. By exploiting the community affiliations of user acco
unts, an attacker may infer sensitive user attributes. This raises the problem o
f community structure deception (CSD), which asks for ways to minimally modify t
he network so that a given community structure maximally hides itself from commu
nity detection algorithms. We investigate CSD through an information-theoretic l
ens. To this end, we propose a community-based structural entropy to express the
 amount of information revealed by a community structure. This notion allows us
to devise residual entropy minimization (REM) as an efficient procedure to solve
 CSD. Experimental results over 9 real-world networks and 6 community detection

algorithms show that REM is very effective in obfuscating the community structure as compared to other benchmark methods.
****
Unsupervised Object Segmentation by Redrawing
Mickaël Chen, Thierry Artières, Ludovic Denoyer
Object segmentation is a crucial problem that is usually solved by using supervised learning approaches over very large datasets composed of both images and corresponding object masks. Since the masks have to be provided at pixel level, building such a dataset for any new domain can be very costly. We present ReDO, a new model able to extract objects from images without any annotation in an unsupervised way. It relies on the idea that it should be possible to change the textures or colors of the objects without changing the overall distribution of the dataset. Following this assumption, our approach is based on an adversarial architecture where the generator is guided by an input sample: given an image, it extracts the object mask, then redraws a new object at the same location. The generator is controlled by a discriminator that ensures that the distribution of generated images is aligned to the original one. We experiment with this method on different datasets and demonstrate the good quality of extracted masks.
****
Unlabeled Data Improves Adversarial Robustness
Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, Percy S. Liang
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.
****
Optimal Stochastic and Online Learning with Individual Iterates
Yunwen Lei, Peng Yang, Ke Tang, Ding-Xuan Zhou
Stochastic composite mirror descent (SCMD) is a simple and efficient method able to capture both geometric and composite structures of optimization problems in machine learning. Existing strategies require to take either an average or a random selection of iterates to achieve optimal convergence rates, which, however, can either destroy the sparsity of solutions or slow down the practical training speed. In this paper, we propose a theoretically sound strategy to select an individual iterate of the vanilla SCMD, which is able to achieve optimal rates for both convex and strongly convex problems in a non-smooth learning setting. This strategy of outputting an individual iterate can preserve the sparsity of solutions which is crucial for a proper interpretation in sparse learning problems. We report experimental comparisons with several baseline methods to show the effectiveness of our method in achieving a fast training speed as well as in outputting sparse solutions.
****
The Implicit Bias of AdaGrad on Separable Data
Qian Qian, Xiaoyuan Qian
We study the implicit bias of AdaGrad on separable linear classification problems.
We show that AdaGrad converges to a direction that can be characterized as the solution of a quadratic optimization problem with the same feasible set as the hard SVM problem.
We also give a discussion about how different choices of the hyperparameters of AdaGrad may impact this direction.
This provides a deeper understanding of why adaptive methods do not seem to have the generalization ability as good as gradient descent does in practice.
****
iSplit LBI: Individualized Partial Ranking with Ties via Split LBI
Qianqian Xu, Xinwei Sun, Zhiyong Yang, Xiaochun Cao, Qingming Huang, Yuan Yao
Due to the inherent uncertainty of data, the problem of predicting partial ranking from pairwise comparison data with ties has attracted increasing interest in recent years. However, in real-world scenarios, different individuals often hold distinct preferences, thus might be misleading to merely look at a global parti

al ranking while ignoring personal diversity. In this paper, instead of learning a global ranking which is agreed with the consensus, we pursue the tie-aware partial ranking from an individualized perspective. Particularly, we formulate a unified framework which not only can be used for individualized partial ranking prediction, but can also be helpful for abnormal users selection. This is realized by a variable splitting-based algorithm called iSplit LBI. Specifically, our algorithm generates a sequence of estimations with a regularization path, where both the hyperparameters and model parameters are updated. At each step of the path, the parameters can be decomposed into three orthogonal parts, namely, abnormal signals, personalized signals and random noise. The abnormal signals can serve the purpose of abnormal user selection, while the abnormal signals and personalized signals
together are mainly responsible for user partial ranking prediction. Extensive experiments on simulated and real-world datasets demonstrate that our new approach significantly outperforms state-of-the-art alternatives.
****
PointDAN: A Multi-Scale 3D Domain Adaption Network for Point Cloud Representation
Can Qin, Haoxuan You, Lichen Wang, C.-C. Jay Kuo, Yun Fu
Domain Adaptation (DA) approaches achieved significant improvements in a wide range of machine learning and computer vision tasks (i.e., classification, detection, and segmentation). However, as far as we are aware, there are few methods yet to achieve domain adaptation directly on 3D point cloud data. The unique challenge of point cloud data lies in its abundant spatial geometric information, and the semantics of the whole object is contributed by including regional geometric structures. Specifically, most general-purpose DA methods that struggle for global feature alignment and ignore local geometric information are not suitable for 3D domain alignment. In this paper, we propose a novel 3D Domain Adaptation Network for point cloud data (PointDAN). PointDAN jointly aligns the global and local features in multi-level. For local alignment, we propose Self-Adaptive (SA) node module with an adjusted receptive field to model the discriminative local structures for aligning domains. To represent hierarchically scaled features, node-attention module is further introduced to weight the relationship of SA nodes across objects and domains. For global alignment, an adversarial-training strategy is employed to learn and align global features across domains. Since there is no common evaluation benchmark for 3D point cloud DA scenario, we build a general benchmark (i.e., PointDA-10) extracted from three popular 3D object/scene datasets (i.e., ModelNet, ShapeNet and ScanNet) for cross-domain 3D objects classification fashion. Extensive experiments on PointDA-10 illustrate the superiority of our model over the state-of-the-art general-purpose DA methods.
****
Certified Adversarial Robustness with Additive Noise
Bai Li, Changyou Chen, Wenlin Wang, Lawrence Carin
The existence of adversarial data examples has drawn significant attention in the deep-learning community; such data are seemingly minimally perturbed relative to the original data, but lead to very different outputs from a deep-learning algorithm. Although a significant body of work on developing defense models has been developed, most such models are heuristic and are often vulnerable to adaptive attacks. Defensive methods that provide theoretical robustness guarantees have been studied intensively, yet most fail to obtain non-trivial robustness when a large-scale model and data are present. To address these limitations, we introduce a framework that is scalable and provides certified bounds on the norm of the input manipulation for constructing adversarial examples. We establish a connection between robustness against adversarial perturbation and additive random noise, and propose a training strategy that can significantly improve the certified bounds. Our evaluation on MNIST, CIFAR-10 and ImageNet suggests that our method is scalable to complicated models and large data sets, while providing competitive robustness to state-of-the-art provable defense methods.
****
Self-Critical Reasoning for Robust Visual Question Answering

Jialin Wu, Raymond Mooney
Visual Question Answering (VQA) deep-learning systems tend to capture superficial statistical correlations in the training data because of strong language priors and fail to generalize to test data with a significantly different question-answer (QA) distribution. To address this issue, we introduce a self-critical training objective that ensures that visual explanations of correct answers match the most influential image regions more than other competitive answer candidates. The influential regions are either determined from human visual/textual explanations or automatically from just significant words in the question and answer. We evaluate our approach on the VQA generalization task using the VQA-CP dataset, achieving a new state-of-the-art i.e. 49.5\% using textual explanations and 48.5\% using automatically
****
Optimal Pricing in Repeated Posted-Price Auctions with Different Patience of the Seller and the Buyer
Arsenii Vanunts, Alexey Drutsa
We study revenue optimization pricing algorithms for repeated posted-price auctions where a seller interacts with a single strategic buyer that holds a fixed private valuation.
When the participants non-equally discount their cumulative utilities, we show that the optimal constant pricing (which offers the Myerson price) is no longer optimal.
In the case of more patient seller, we propose a novel multidimensional optimization functional --- a generalization of the one used to determine Myerson's price. This functional allows to find the optimal algorithm and to boost revenue of the optimal static pricing by an efficient low-dimensional approximation.
Numerical experiments are provided to support our results.
****
Stand-Alone Self-Attention in Vision Models
Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, Jon Shlens
Convolutions are a fundamental building block of modern computer vision systems. Recent approaches have argued for going beyond convolutions in order to capture long-range dependencies. These efforts focus on augmenting convolutional models with content-based interactions, such as self-attention and non-local means, to achieve gains on a number of vision tasks. The natural question that arises is whether attention can be a stand-alone primitive for vision models instead of serving as just an augmentation on top of convolutions. In developing and testing a pure self-attention vision model, we verify that self-attention can indeed be an effective stand-alone layer. A simple procedure of replacing all instances of spatial convolutions with a form of self-attention to ResNet-50 produces a fully self-attentional model that outperforms the baseline on ImageNet classification with 12% fewer FLOPS and 29% fewer parameters. On COCO object detection, a fully self-attention model matches the mAP of a baseline RetinaNet while having 39% fewer FLOPS and 34% fewer parameters. Detailed ablation studies demonstrate that self-attention is especially impactful when used in later layers. These results establish that stand-alone self-attention is an important addition to the vision practitioner's toolbox.
****
Debiased Bayesian inference for average treatment effects
Kolyan Ray, Botond Szabo
Bayesian approaches have become increasingly popular in causal inference problems due to their conceptual simplicity, excellent performance and in-built uncertainty quantification ('posterior credible sets'). We investigate Bayesian inference for average treatment effects from observational data, which is a challenging problem due to the missing counterfactuals and selection bias. Working in the standard potential outcomes framework, we propose a data-driven modification to an arbitrary (nonparametric) prior based on the propensity score that corrects for the first-order posterior bias, thereby improving performance. We illustrate our method for Gaussian process (GP) priors using (semi-)synthetic data. Our expe

riments demonstrate significant improvement in both estimation accuracy and uncertainty quantification compared to the unmodified GP, rendering our approach highly competitive with the state-of-the-art.
****
Globally optimal score-based learning of directed acyclic graphs in high-dimensions
Bryon Aragam, Arash Amini, Qing 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.
****
GIFT: Learning Transformation-Invariant Dense Visual Descriptors via Group CNNs
Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, Xiaowei Zhou
Finding local correspondences between images with different viewpoints requires local descriptors that are robust against geometric transformations. An approach for transformation invariance is to integrate out the transformations by pooling the features extracted from transformed versions of an image. However, the feature pooling may sacrifice the distinctiveness of the resulting descriptors. In this paper, we introduce a novel visual descriptor named Group Invariant Feature Transform (GIFT), which is both discriminative and robust to geometric transformations. The key idea is that the features extracted from the transformed versions of an image can be viewed as a function defined on the group of the transformations. Instead of feature pooling, we use group convolutions to exploit underlying structures of the extracted features on the group, resulting in descriptors that are both discriminative and provably invariant to the group of transformations. Extensive experiments show that GIFT outperforms state-of-the-art methods on several benchmark datasets and practically improves the performance of relative pose estimation.
****
Convergence of Adversarial Training in Overparametrized Neural Networks
Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, 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.
****
Explicit Disentanglement of Appearance and Perspective in Generative Models
Nicki Skafte, Søren Hauberg
Disentangled representation learning finds compact, independent and easy-to-interpret factors of the data.
Learning such has been shown to require an inductive bias, which we explicitly encode in a generative model of images. Specifically, we propose a model with two latent spaces: one that represents spatial transformations of the input data, and another that represents the transformed data. We find that the latter naturally captures the intrinsic appearance of the data. To realize the generative model, we propose a Variationally Inferred Transformational Autoencoder (VITAE) that incorporates a spatial ransformer into a variational autoencoder. We show how to perform inference in the model efficiently by carefully designing the encoders and restricting the transformation class to be diffeomorphic. Empirically, our model separates the visual style from digit type on MNIST, separates shape and pose in images of human bodies and facial features from facial shape on CelebA.
****
Fast and Furious Learning in Zero-Sum Games: Vanishing Regret with Non-Vanishing Step Sizes
James Bailey, Georgios Piliouras
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.

****
Slice-based Learning: A Programming Model for Residual Learning in Critical Data Slices
Vincent Chen, Sen Wu, Alexander J. Ratner, Jen Weng, Christopher Ré
In real-world machine learning applications, data subsets correspond to especially critical outcomes: vulnerable cyclist detections are safety-critical in an autonomous driving task, and "question" sentences might be important to a dialogue agent's language understanding for product purposes. While machine learning models can achieve quality performance on coarse-grained metrics like F1-score and overall accuracy, they may underperform on these critical subsets---we define these as slices, the key abstraction in our approach. To address slice-level performance, practitioners often train separate "expert" models on slice subsets or use multi-task hard parameter sharing. We propose Slice-based Learning, a new programming model in which the slicing function (SF), a programmer abstraction, is used to specify additional model capacity for each slice. Any model can leverage SFs to learn slice-specific representations, which are combined with an attention mechanism to make slice-aware predictions. We show that our approach improves over baselines in terms of computational complexity and slice-specific performance by up to 19.0 points, and overall performance by up to 4.6 F1 points on applications spanning natural language understanding and computer vision benchmarks as well as production-scale industrial systems.
****
Nearly Tight Bounds for Robust Proper Learning of Halfspaces with a Margin
Ilias Diakonikolas, Daniel Kane, Pasin Manurangsi
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.
****
Distribution-Independent PAC Learning of Halfspaces with Massart Noise
Ilias Diakonikolas, Themis Gouleakis, Christos Tzamos
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.
****
Poisson-Minibatching for Gibbs Sampling with Convergence Rate Guarantees
Ruqi Zhang, Christopher M. De Sa
Gibbs sampling is a Markov chain Monte Carlo method that is often used for learning and inference on graphical models.
Minibatching, in which a small random subset of the graph is used at each iteration, can help make Gibbs sampling scale to large graphical models by reducing its computational cost.
In this paper, we propose a new auxiliary-variable minibatched Gibbs sampling method, {\it Poisson-minibatching Gibbs}, which both produces unbiased samples and has a theoretical guarantee on its convergence rate.
In comparison to previous minibatched Gibbs algorithms, Poisson-minibatching Gibbs supports fast sampling from continuous state spaces and avoids the need for a Metropolis-Hastings correction on discrete state spaces.
We demonstrate the effectiveness of our method on multiple applications and in comparison with both plain Gibbs and previous minibatched methods.
****
Semi-Parametric Dynamic Contextual Pricing
Virag Shah, Ramesh Johari, Jose Blanchet
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.
****
Theoretical evidence for adversarial robustness through randomization

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cedric Gouy-Pailler, Jamal Atif
This paper investigates the theory of robustness against adversarial attacks. It focuses on the family of randomization techniques that consist in injecting nois
e
in the network at inference time. These techniques have proven effective in many contexts, but lack theoretical arguments. We close this gap by presenting a theo
-
retical analysis of these approaches, hence explaining why they perform well in practice. More precisely, we make two new contributions. The first one relates the randomization rate to robustness to adversarial attacks. This result applies for
the general family of exponential distributions, and thus extends and unifies th
e
previous approaches. The second contribution consists in devising a new upper bound on the adversarial risk gap of randomized neural networks. We support our theoretical claims with a set of experiments.
****
On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks
Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, Sarah Michalak
Mixup~\cite{zhang2017mixup} is a recently proposed method for training deep neural networks where additional samples are generated during training by convexly combining random pairs of images and their associated labels. While simple to implement, it has shown to be a surprisingly effective method of data augmentation for image classification; DNNs trained with mixup show noticeable gains in classification performance on a number of image classification benchmarks. In this work, we discuss a hitherto untouched aspect of mixup training -- the calibration and predictive uncertainty   of models trained with mixup. We find that DNNs trained with mixup are significantly better calibrated -- i.e the predicted softmax scores are much better indicators of the actual likelihood of a correct prediction -- than DNNs trained in the regular fashion. We conduct experiments on a number of image classification architectures and datasets -- including large-scale datasets like ImageNet -- and find this to be the case.
    Additionally, we find that merely mixing features does not result in the same calibration benefit and that the label smoothing in mixup training plays a significant role in improving calibration. Finally, we also observe that mixup-trained DNNs are less prone to over-confident predictions on out-of-distribution and random-noise data. We conclude that the typical overconfidence seen in neural networks, even on in-distribution data is likely a consequence of training with hard labels, suggesting that mixup training be employed for classification tasks where predictive uncertainty is a significant concern.
****
Thompson Sampling for Multinomial Logit Contextual Bandits
Min-hwan Oh, Garud Iyengar
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.
****
Symmetry-Based Disentangled Representation Learning requires Interaction with Environments
Hugo Caselles-Dupré, Michael Garcia Ortiz, David Filliat
Finding a generally accepted formal definition of a disentangled representation in the context of an agent behaving in an environment is an important challenge towards the construction of data-efficient autonomous agents. Higgins et al. recently proposed Symmetry-Based Disentangled Representation Learning, a definition based on a characterization of symmetries in the environment using group theory. We build on their work and make observations, theoretical and empirical, that

lead us to argue that Symmetry-Based Disentangled Representation Learning cannot only be based on static observations: agents should interact with the environment to discover its symmetries. Our experiments can be reproduced in Colab and the code is available on GitHub.
****
Mining GOLD Samples for Conditional GANs
Sangwoo Mo, Chiheon Kim, Sungwoong Kim, Minsu Cho, Jinwoo Shin
Conditional generative adversarial networks (cGANs) have gained a considerable attention in recent years due to its class-wise controllability and superior quality for complex generation tasks. We introduce a simple yet effective approach to improving cGANs by measuring the discrepancy between the data distribution and the model distribution on given samples. The proposed measure, coined the gap of log-densities (GOLD), provides an effective self-diagnosis for cGANs while being efficiently, computed from the discriminator. We propose three applications of the GOLD: example re-weighting, rejection sampling, and active learning, which improve the training, inference, and data selection of cGANs, respectively. Our experimental results demonstrate that the proposed methods outperform corresponding baselines for all three applications on different image datasets.
****
Few-shot Video-to-Video Synthesis
Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, Jan Kautz
Video-to-video synthesis (vid2vid) aims at converting an input semantic video, such as videos of human poses or segmentation masks, to an output photorealistic video. While the state-of-the-art of vid2vid has advanced significantly, existing approaches share two major limitations. First, they are data-hungry. Numerous images of a target human subject or a scene are required for training. Second, a learned model has limited generalization capability. A pose-to-human vid2vid model can only synthesize poses of the single person in the training set. It does not generalize to other humans that are not in the training set. To address the limitations, we propose a few-shot vid2vid framework, which learns to synthesize videos of previously unseen subjects or scenes by leveraging few example images of the target at test time. Our model achieves this few-shot generalization capability via a novel network weight generation module utilizing an attention mechanism. We conduct extensive experimental validations with comparisons to strong baselines using several large-scale video datasets including human-dancing videos, talking-head videos, and street-scene videos. The experimental results verify the effectiveness of the proposed framework in addressing the two limitations of existing vid2vid approaches.
****
Unlocking Fairness: a Trade-off Revisited
Michael Wick, swetasudha panda, Jean-Baptiste Tristan
The prevailing wisdom is that a model's fairness and its accuracy are in tension with one another. However, there is a pernicious {\em modeling-evaluating dualism} bedeviling fair machine learning in which phenomena such as label bias are appropriately acknowledged as a source of unfairness when designing fair models, only to be tacitly abandoned when evaluating them. We investigate fairness and accuracy, but this time under a variety of controlled conditions in which we vary the amount and type of bias. We find, under reasonable assumptions, that the tension between fairness and accuracy is illusive, and vanishes as soon as we account for these phenomena during evaluation. Moreover, our results are consistent with an opposing conclusion: fairness and accuracy are sometimes in accord. This raises the question, {\em might there be a way to harness fairness to improve accuracy after all?} Since most notions of fairness are with respect to the model's predictions and not the ground truth labels, this provides an opportunity to see if we can improve accuracy by harnessing appropriate notions of fairness over large quantities of {\em unlabeled} data with techniques like posterior regularization and generalized

expectation. Indeed, we find that semi-supervision not only
improves fairness, but also accuracy and has advantages over
existing in-processing methods that succumb to selection bias on the
training set.
****
Stochastic Shared Embeddings: Data-driven Regularization of Embedding Layers
Liwei Wu, Shuqing Li, Cho-Jui Hsieh, James L. Sharpnack
In deep neural nets, lower level embedding layers account for a large portion of the total number of parameters. Tikhonov regularization, graph-based regularization, and hard parameter sharing are approaches that introduce explicit biases into training in a hope to reduce statistical complexity. Alternatively, we propose stochastically shared embeddings (SSE), a data-driven approach to regularizing embedding layers, which stochastically transitions between embeddings during stochastic gradient descent (SGD). Because SSE integrates seamlessly with existing SGD algorithms, it can be used with only minor modifications when training large scale neural networks. We develop two versions of SSE: SSE-Graph using knowledge graphs of embeddings; SSE-SE using no prior information. We provide theoretical guarantees for our method and show its empirical effectiveness on 6 distinct tasks, from simple neural networks with one hidden layer in recommender systems, to the transformer and BERT in natural languages. We find that when used along with widely-used regularization methods such as weight decay and dropout, our proposed SSE can further reduce overfitting, which often leads to more favorable generalization results.
****
An Algorithmic Framework For Differentially Private Data Analysis on Trusted Processors
Joshua Allen, Bolin Ding, Janardhan Kulkarni, Harsha Nori, Olga Ohrimenko, Sergey Yekhanin
Differential privacy has emerged as the main definition for private data analysis and machine learning. The global model of differential privacy, which assumes that users trust the data collector, provides strong privacy guarantees and introduces small errors in the output. In contrast, applications of differential privacy in commercial systems by Apple, Google, and Microsoft, use the local model. Here, users do not trust the data collector, and hence randomize their data before sending it to the data collector. Unfortunately, local model is too strong for several important applications and hence is limited in its applicability. In this work, we propose a framework based on trusted processors and a new definition of differential privacy called Oblivious Differential Privacy, which combines the best of both local and global models. The algorithms we design in this framework show interesting interplay of ideas from the streaming algorithms, oblivious algorithms, and differential privacy.
****
Implicit Generation and Modeling with Energy Based Models
Yilun Du, Igor Mordatch
Energy based models (EBMs) are appealing due to their generality and simplicity in likelihood modeling, but have been traditionally difficult to train. We present techniques to scale MCMC based EBM training on continuous neural networks, and we show its success on the high-dimensional data domains of ImageNet32x32, ImageNet128x128, CIFAR-10, and robotic hand trajectories, achieving better samples than other likelihood models and nearing the performance of contemporary GAN approaches, while covering all modes of the data. We highlight some unique capabilities of implicit generation such as compositionality and corrupt image reconstruction and inpainting. Finally, we show that EBMs are useful models across a wide variety of tasks, achieving state-of-the-art out-of-distribution classification, adversarially robust classification, state-of-the-art continual online class learning, and coherent long term predicted trajectory rollouts.
****
Evaluating Protein Transfer Learning with TAPE
Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, Yun Song

Protein modeling is an increasingly popular area of machine learning research. Semi-supervised learning has emerged as an important paradigm in protein modeling due to the high cost of acquiring supervised protein labels, but the current literature is fragmented when it comes to datasets and standardized evaluation techniques. To facilitate progress in this field, we introduce the Tasks Assessing Protein Embeddings (TAPE), a set of five biologically relevant semi-supervised learning tasks spread across different domains of protein biology. We curate tasks into specific training, validation, and test splits to ensure that each task tests biologically relevant generalization that transfers to real-life scenarios. We benchmark a range of approaches to semi-supervised protein representation learning, which span recent work as well as canonical sequence learning techniques. We find that self-supervised pretraining is helpful for almost all models on all tasks, more than doubling performance in some cases. Despite this increase, in several cases features learned by self-supervised pretraining still lag behind features extracted by state-of-the-art non-neural techniques. This gap in performance suggests a huge opportunity for innovative architecture design and improved modeling paradigms that better capture the signal in biological sequences. TAPE will help the machine learning community focus effort on scientifically relevant problems. Toward this end, all data and code used to run these experiments is available at https://github.com/songlab-cal/tape
****
Recurrent Space-time Graph Neural Networks
Andrei Nicolicioiu, Iulia Duta, Marius Leordeanu
Learning in the space-time domain remains a very challenging problem in machine learning and computer vision. Current computational models for understanding spatio-temporal visual data are heavily rooted in the classical single-image based paradigm. It is not yet well understood how to integrate information in space and time into a single, general model. We propose a neural graph model, recurrent in space and time, suitable for capturing both the local appearance and the complex higher-level interactions of different entities and objects within the changing world scene. Nodes and edges in our graph have dedicated neural networks for processing information. Nodes operate over features extracted from local parts in space and time and over previous memory states. Edges process messages between connected nodes at different locations and spatial scales or between past and present time. Messages are passed iteratively in order to transmit information globally and establish long range interactions. Our model is general and could learn to recognize a variety of high level spatio-temporal concepts and be applied to different learning tasks. We demonstrate, through extensive experiments and ablation studies, that our model outperforms strong baselines and top published methods on recognizing complex activities in video. Moreover, we obtain state-of-the-art performance on the challenging Something-Something human-object interaction dataset.
****
Singleshot : a scalable Tucker tensor decomposition
Abraham Traore, Maxime Berar, Alain Rakotomamonjy
This paper introduces a new approach for the scalable Tucker decomposition problem. Given a tensor X , the method proposed allows to infer the latent factors
by processing one subtensor drawn from X at a time. The key principle of our approach is based on the recursive computations of gradient and on cyclic update of factors involving only one single step of gradient descent. We further improve the
computational efficiency of this algorithm by proposing an inexact gradient version.
These two algorithms are backed with theoretical guarantees of convergence and convergence rate under mild conditions. The scalabilty of the proposed approaches
which can be easily extended to handle some common constraints encountered in tensor decomposition (e.g non-negativity), is proven via numerical experiments on

both synthetic and real data sets.
****
Secretary Ranking with Minimal Inversions
Sepehr Assadi, Eric Balkanski, Renato Leme
We study a secretary problem which captures the task of ranking in online settings. We term this problem the secretary ranking problem: elements from an ordered set arrive in random order and instead of picking the maximum element, the algorithm is asked to assign a rank, or position, to each of the elements. The rank assigned is irrevocable and is given knowing only the pairwise comparisons with elements previously arrived. The goal is to minimize the distance of the rank produced to the true rank of the elements measured by the Kendall-Tau distance, which corresponds to the number of pairs that are inverted with respect to the true order.
****
Policy Continuation with Hindsight Inverse Dynamics
Hao Sun, Zhizhong Li, Xiaotong Liu, Bolei Zhou, Dahua Lin
Solving goal-oriented tasks is an important but challenging problem in reinforcement learning (RL). For such tasks, the rewards are often sparse, making it difficult to learn a policy effectively. To tackle this difficulty, we propose a new approach called Policy Continuation with Hindsight Inverse Dynamics (PCHID). This approach learns from Hindsight Inverse Dynamics based on Hindsight Experience Replay. Enabling the learning process in a self-imitated manner and thus can be trained with supervised learning. This work also extends it to multi-step settings with Policy Continuation. The proposed method is general, which can work in isolation or be combined with other on-policy and off-policy algorithms. On two multi-goal tasks GridWorld and FetchReach, PCHID significantly improves the sample efficiency as well as the final performance.
****
A Mean Field Theory of Quantized Deep Networks: The Quantization-Depth Trade-Off
Yaniv Blumenfeld, Dar Gilboa, 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.
****
Exponentially convergent stochastic k-PCA without variance reduction
Cheng Tang
We present Matrix Krasulina, an algorithm for online k-PCA, by gen- eralizing the classic Krasulina’s method (Krasulina, 1969) from vector to matrix case. We show, both theoretically and empirically, that the algorithm naturally adapts to data low-rankness and converges exponentially fast to the ground-truth principal subspace. Notably, our result suggests that despite various recent efforts to accelerate the convergence of stochastic-gradient based methods by adding a O(n)-time variance reduction step, for the k- PCA problem, a truly online SGD variant suffices to achieve exponential convergence on intrinsically low-rank data.
****
DISN: Deep Implicit Surface Network for High-quality Single-view 3D Reconstruction
Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, Ulrich Neumann
Reconstructing 3D shapes from single-view images has been a long-standing
research problem. In this paper, we present DISN, a Deep Implicit Surface Net-
work which can generate a high-quality detail-rich 3D mesh from a 2D image by
predicting the underlying signed distance fields. In addition to utilizing global
image features, DISN predicts the projected location for each 3D point on the
2D image and extracts local features from the image feature maps. Combin-
ing global and local features significantly improves the accuracy of the signed
distance field prediction, especially for the detail-rich areas. To the best of our
knowledge, DISN is the first method that constantly captures details such as

holes and thin structures present in 3D shapes from single-view images. DISN achieves the state-of-the-art single-view reconstruction performance on a variet
y
of shape categories reconstructed from both synthetic and real images. Code is available at https://github.com/laughtervv/DISN. The supplemen-
tary can be found at https://xharlie.github.io/images/neurips_
2019_supp.pdf
****
Personalizing Many Decisions with High-Dimensional Covariates
Nima Hamidi, Mohsen Bayati, Kapil Gupta
We consider the k-armed stochastic contextual bandit problem with d dimensional features, when both k and d can be large. To the best of our knowledge, all exis ting algorithm for this problem have a regret bound that scale as polynomials of degree at least two in k and d. The main contribution of this paper is to intro duce and theoretically analyze a new algorithm (REAL Bandit) with a regret that scales by r^2(k+d) when r is rank of the k by d matrix of unknown parameters. RE AL Bandit relies on ideas from low-rank matrix estimation literature and a new r ow-enhancement subroutine that yields sharper bounds for estimating each row of the parameter matrix that may be of independent interest.
****
Universal Approximation of Input-Output Maps by Temporal Convolutional Nets
Joshua Hanson, Maxim Raginsky
There has been a recent shift in sequence-to-sequence modeling from recurrent ne twork architectures to convolutional network architectures due to computational advantages in training and operation while still achieving competitive performa nce. For systems having limited long-term temporal dependencies, the approximati on capability of recurrent networks is essentially equivalent to that of tempora l convolutional nets (TCNs). We prove that TCNs can approximate a large class of input-output maps having approximately finite memory to arbitrary error toleran ce. Furthermore, we derive quantitative approximation rates for deep ReLU TCNs i n terms of the width and depth of the network and modulus of continuity of the o riginal input-output map, and apply these results to input-output maps of system s that admit finite-dimensional state-space realizations (i.e., recurrent models).
****
Equipping Experts/Bandits with Long-term Memory
Kai Zheng, Haipeng Luo, Ilias Diakonikolas, Liwei 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Function-Space Distributions over Kernels
Gregory Benton, Wesley J. Maddox, Jayson Salkey, Julio Albinati, Andrew Gordon W ilson
Gaussian processes are flexible function approximators, with inductive biases co ntrolled by a covariance kernel. Learning the kernel is the key to representatio n learning and strong predictive performance. In this paper, we develop function al kernel learning (FKL) to directly infer functional posteriors over kernels. I n particular, we place a transformed Gaussian process over a spectral density, t o induce a non-parametric distribution over kernel functions. The resulting appr oach enables learning of rich representations, with support for any stationary k ernel, uncertainty over the values of the kernel, and an interpretable specifica tion of a prior directly over kernels, without requiring sophisticated initializ ation or manual intervention. We perform inference through elliptical slice samp ling, which is especially well suited to marginalizing posteriors with the stron gly correlated priors typical to function space modeling. We develop our approac h for non-uniform, large-scale, multi-task, and multidimensional data, and show promising performance in a wide range of settings, including interpolation, extr apolation, and kernel recovery experiments.

****
Fully Neural Network based Model for General Temporal Point Processes
Takahiro Omi, naonori ueda, Kazuyuki Aihara
A temporal point process is a mathematical model for a time series of discrete events, which covers various applications. Recently, recurrent neural network (RNN) based models have been developed for point processes and have been found effective. RNN based models usually assume a specific functional form for the time course of the intensity function of a point process (e.g., exponentially decreasing or increasing with the time since the most recent event). However, such an assumption can restrict the expressive power of the model. We herein propose a novel RNN based model in which the time course of the intensity function is represented in a general manner. In our approach, we first model the integral of the intensity function using a feedforward neural network and then obtain the intensity function as its derivative. This approach enables us to both obtain a flexible model of the intensity function and exactly evaluate the log-likelihood function, which contains the integral of the intensity function, without any numerical approximations. Our model achieves competitive or superior performances compared to the previous state-of-the-art methods for both synthetic and real datasets.
****
Splitting Steepest Descent for Growing Neural Architectures
Lemeng Wu, Dilin Wang, Qiang 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.
****
Improving Textual Network Learning with Variational Homophilic Embeddings
Wenlin Wang, Chenyang Tao, Zhe Gan, Guoyin Wang, Liqun Chen, Xinyuan Zhang, Ruiyi Zhang, Qian Yang, Ricardo Henao, Lawrence Carin
The performance of many network learning applications crucially hinges on the success of network embedding algorithms, which aim to encode rich network information into low-dimensional vertex-based vector representations. This paper considers a novel variational formulation of network embeddings, with special focus on textual networks. Different from most existing methods that optimize a discriminative objective, we introduce Variational Homophilic Embedding (VHE), a fully generative model that learns network embeddings by modeling the semantic (textual) information with a variational autoencoder, while accounting for the structural (topology) information through a novel homophilic prior design. Homophilic vertex embeddings encourage similar embedding vectors for related (connected) vertices. The VHE encourages better generalization for downstream tasks, robustness to incomplete observations, and the ability to generalize to unseen vertices. Extensive experiments on real-world networks, for multiple tasks, demonstrate that the proposed method achieves consistently superior performance relative to competing state-of-the-art approaches.
****
Deep Supervised Summarization: Algorithm and Application to Learning Instructions
Chengguang Xu, Ehsan Elhamifar
We address the problem of finding representative points of datasets by learning from multiple datasets and their ground-truth summaries. We develop a supervised subset selection framework, based on the facility location utility function, which learns to map datasets to their ground-truth representatives. To do so, we propose to learn representations of data so that the input of transformed data to the facility location recovers their ground-truth representatives. Given the NP-hardness of the utility function, we consider its convex relaxation based on sparse representation and investigate conditions under which the solution of the convex optimization recovers ground-truth representatives of each dataset. We design a loss function whose minimization over the parameters of the data representation network leads to satisfying the theoretical conditions, hence guaranteeing recovering ground-truth summaries. Given the non-convexity of the loss function

, we develop an efficient learning scheme that alternates between representation learning by minimizing our proposed loss given the current assignments of points to ground-truth representatives and updating assignments given the current data representation. By experiments on the problem of learning key-steps (subactivities) of instructional videos, we show that our proposed framework improves the state-of-the-art supervised subset selection algorithms.
****
Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting
Rajat Sen, Hsiang-Fu Yu, Inderjit S. Dhillon
Forecasting high-dimensional time series plays a crucial role in many applications such as demand forecasting and financial predictions. Modern datasets can have millions of correlated time-series that evolve together, i.e they are extremely high dimensional (one dimension for each individual time-series). There is a need for exploiting global patterns and coupling them with local calibration for better prediction. However, most recent deep learning approaches in the literature are one-dimensional, i.e, even though they are trained on the whole dataset, during prediction, the future forecast for a single dimension mainly depends on past values from the same dimension. In this paper, we seek to correct this deficiency and propose DeepGLO, a deep forecasting model which thinks globally and acts locally. In particular, DeepGLO is a hybrid model that combines a global matrix factorization model regularized by a temporal convolution network, along with another temporal network that can capture local properties of each time-series and associated covariates. Our model can be trained effectively on high-dimensional but diverse time series, where different time series can have vastly different scales, without a priori normalization or rescaling. Empirical results demonstrate that DeepGLO can outperform state-of-the-art approaches; for example, we see more than 25% improvement in WAPE over other methods on a public dataset that contains more than 100K-dimensional time series.
****
Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso
Quentin Bertrand, Mathurin Massias, Alexandre Gramfort, Joseph Salmon
A limitation of Lasso-type estimators is that the optimal regularization parameter depends on the unknown noise level. Estimators such as the concomitant Lasso address this dependence by jointly estimating the noise level and the regression coefficients. Additionally, in many applications, the data is obtained by averaging multiple measurements: this reduces the noise variance, but it dramatically reduces sample sizes and prevents refined noise modeling. In this work, we propose a concomitant estimator that can cope with complex noise structure by using non-averaged measurements, its data-fitting term arising as a smoothing of the nuclear norm. The resulting optimization problem is convex and amenable, thanks to smoothing theory, to state-of-the-art optimization techniques that leverage the sparsity of the solutions. Practical benefits are demonstrated on toy datasets, realistic simulated data and real neuroimaging data.
****
PAC-Bayes under potentially heavy tails
Matthew Holland
We derive PAC-Bayesian learning guarantees for heavy-tailed losses, and obtain a novel optimal Gibbs posterior which enjoys finite-sample excess risk bounds at logarithmic confidence. Our core technique itself makes use of PAC-Bayesian inequalities in order to derive a robust risk estimator, which by design is easy to compute. In particular, only assuming that the first three moments of the loss distribution are bounded, the learning algorithm derived from this estimator achieves nearly sub-Gaussian statistical error, up to the quality of the prior.
****
Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers
Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, Greg 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Regression Planning Networks
Danfei Xu, Roberto Martín-Martín, De-An Huang, Yuke Zhu, Silvio Savarese, Li F. Fei-Fei
Recent learning-to-plan methods have shown promising results on planning directl y from observation space. Yet, their ability to plan for long-horizon tasks is l imited by the accuracy of the prediction model. On the other hand, classical sym bolic planners show remarkable capabilities in solving long-horizon tasks, but t hey require predefined symbolic rules and symbolic states, restricting their rea l-world applicability. In this work, we combine the benefits of these two paradi gms and propose a learning-to-plan method that can directly generate a long-term symbolic plan conditioned on high-dimensional observations. We borrow the idea of regression (backward) planning from classical planning literature and introdu ce Regression Planning Networks (RPN), a neural network architecture that plans backward starting at a task goal and generates a sequence of intermediate goals that reaches the current observation. We show that our model not only inherits m any favorable traits from symbolic planning --including the ability to solve pre viously unseen tasks-- but also can learn from visual inputs in an end-to-end ma nner. We evaluate the capabilities of RPN in a grid world environment and a simu lated 3D kitchen environment featuring complex visual scenes and long task horiz on, and show that it achieves near-optimal performance in completely new task in stances.
****
Efficient Neural Architecture Transformation Search in Channel-Level for Object Detection
Junran Peng, Ming Sun, ZHAO-XIANG ZHANG, Tieniu Tan, Junjie Yan
Recently, Neural Architecture Search has achieved great success in large-scale i mage classification. In contrast, there have been limited works focusing on arch itecture search for object detection, mainly because the costly ImageNet pretrai ning is always required for detectors. Training from scratch, as a substitute, d emands more epochs to converge and brings no computation saving.
****
CXPlain: Causal Explanations for Model Interpretation under Uncertainty
Patrick Schwab, Walter Karlen
Feature importance estimates that inform users about the degree to which given i nputs influence the output of a predictive model are crucial for understanding, validating, and interpreting machine-learning models. However, providing fast an d accurate estimates of feature importance for high-dimensional data, and quanti fying the uncertainty of such estimates remain open challenges. Here, we frame t he task of providing explanations for the decisions of machine-learning models a s a causal learning task, and train causal explanation (CXPlain) models that lea rn to estimate to what degree certain inputs cause outputs in another machine-le arning model. CXPlain can, once trained, be used to explain the target model in little time, and enables the quantification of the uncertainty associated with i ts feature importance estimates via bootstrap ensembling. We present experiments that demonstrate that CXPlain is significantly more accurate and faster than ex isting model-agnostic methods for estimating feature importance. In addition, we confirm that the uncertainty estimates provided by CXPlain ensembles are strong ly correlated with their ability to accurately estimate feature importance on he ld-out data.
****
Compacting, Picking and Growing for Unforgetting Continual Learning
Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, Chu-Son g Chen
Continual lifelong learning is essential to many applications. In this paper, we propose a simple but effective approach to continual deep learning. Our approac h leverages the principles of deep model compression, critical weights selection

, and progressive networks expansion. By enforcing their integration in an itera tive manner, we introduce an incremental learning method that is scalable to the number of sequential tasks in a continual learning process. Our approach is eas y to implement and owns several favorable characteristics. First, it can avoid f orgetting (i.e., learn new tasks while remembering all previous tasks). Second, it allows model expansion but can maintain the model compactness when handling s equential tasks. Besides, through our compaction and selection/expansion mechani sm, we show that the knowledge accumulated through learning previous tasks is he lpful to build a better model for the new tasks compared to training the models independently with tasks. Experimental results show that our approach can increm entally learn a deep model tackling multiple tasks without forgetting, while the model compactness is maintained with the performance more satisfiable than indi vidual task training.
****
Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments
Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, Greg Lewis
We consider the estimation of heterogeneous treatment effects with arbitrary mac hine learning methods in the presence of unobserved confounders with the aid of a valid instrument. Such settings arise in A/B tests with an intent-to-treat str ucture, where the experimenter randomizes over which user will receive a recomme ndation to take an action, and we are interested in the effect of the downstream action. We develop a statistical learning approach to the estimation of heterog eneous effects, reducing the problem to the minimization of an appropriate loss function that depends on a set of auxiliary models (each corresponding to a sepa rate prediction task). The reduction enables the use of all recent algorithmic a dvances (e.g. neural nets, forests). We show that the estimated effect model is robust to estimation errors in the auxiliary models, by showing that the loss sa tisfies a Neyman orthogonality criterion. Our approach can be used to estimate p rojections of the true effect model on simpler hypothesis spaces. When these spa ces are parametric, then the parameter estimates are asymptotically normal, whic h enables construction of confidence sets. We applied our method to estimate the effect of membership on downstream webpage engagement for a major travel webpag e, using as an instrument an intent-to-treat A/B test among 4 million users, whe re some users received an easier membership sign-up process. We also validate ou r method on synthetic data and on public datasets for the effects of schooling o n income.
****
Mapping State Space using Landmarks for Universal Goal Reaching
Zhiao Huang, Fangchen Liu, Hao Su
An agent that has well understood the environment should be able to apply its sk ills for any given goals, leading to the fundamental problem of learning the Uni versal Value Function Approximator (UVFA). A UVFA learns to predict the cumulati ve rewards between all state-goal pairs. However, empirically, the value functio n for long-range goals is always hard to estimate and may consequently result in failed policy. This has presented challenges to the learning process and the ca pability of neural networks. We propose a method to address this issue in large MDPs with sparse rewards, in which exploration and routing across remote states are both extremely challenging. Our method explicitly models the environment in a hierarchical manner, with a high-level dynamic landmark-based map abstracting the visited state space, and a low-level value network to derive precise local d ecisions. We use farthest point sampling to select landmark states from past exp erience, which has improved exploration compared with simple uniform sampling. E xperimentally we showed that our method enables the agent to reach long-range go als at the early training stage, and achieve better performance than standard RL algorithms for a number of challenging tasks.
****
Convergence-Rate-Matching Discretization of Accelerated Optimization Flows Throu gh Opportunistic State-Triggered Control
Miguel Vaquero, Jorge Cortes

A recent body of exciting work seeks to shed light on the behavior of accelerated methods in optimization via high-resolution differential equations. These differential equations are continuous counterparts of the discrete-time optimization algorithms, and their convergence properties can be characterized using the powerful tools provided by classical Lyapunov stability analysis. An outstanding question of pivotal importance is how to discretize these continuous flows while maintaining their convergence rates. This paper provides a novel approach through the idea of opportunistic state-triggered control. We take advantage of the Lyapunov functions employed to characterize the rate of convergence of high-resolution differential equations to design variable-stepsize forward-Euler discretizations that preserve the Lyapunov decay of the original dynamics. The philosophy of our approach is not limited to forward-Euler discretizations and may be combined with other integration schemes.
****
Principal Component Projection and Regression in Nearly Linear Time through Asymmetric SVRG
Yujia Jin, Aaron Sidford
Given a n-by-d data matrix A, principal component projection (PCP) and principal component regression (PCR), i.e. projection and regression restricted to the top-eigenspace of A, are fundamental problems in machine learning, optimization, and numerical analysis. In this paper we provide the first algorithms that solve these problems in nearly linear time for fixed eigenvalue distribution and large n. This improves upon previous methods which had superlinear running times when either the number of top eigenvalues or gap between the eigenspaces were large.

****
Private Stochastic Convex Optimization with Optimal Rates
Raef Bassily, Vitaly Feldman, Kunal Talwar, 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.
****
Complexity of Highly Parallel Non-Smooth Convex Optimization
Sebastien Bubeck, Qijia Jiang, Yin-Tat Lee, Yuanzhi Li, 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.
****
A Structured Prediction Approach for Generalization in Cooperative Multi-Agent Reinforcement Learning
Nicolas Carion, Nicolas Usunier, Gabriel Synnaeve, Alessandro Lazaric
Effective coordination is crucial to solve multi-agent collaborative (MAC) problems. While centralized reinforcement learning methods can optimally solve small MAC instances, they do not scale to large problems and they fail to generalize to scenarios different from those seen during training.
In this paper, we consider MAC problems with some intrinsic notion of locality (e.g., geographic proximity) such that interactions between agents and tasks are locally limited. By leveraging this property, we introduce a novel structured prediction approach to assign agents to tasks. At each step, the assignment is obtained by solving a centralized optimization problem (the inference procedure) whose objective function is parameterized by a learned scoring model. We propose different combinations of inference procedures and scoring models able to represent coordination patterns of increasing complexity. The resulting assignment policy can be efficiently learned on small problem instances and readily reused in problems with more agents and tasks (i.e., zero-shot generalization). We report experimental results on a toy search and rescue problem and on several target selection scenarios in StarCraft: Brood War, in which our model significantly outperforms strong rule-based baselines on instances with 5 times more agents and tas

ks than those seen during training.
****
Interaction Hard Thresholding: Consistent Sparse Quadratic Regression in Sub-quadratic Time and Space
Shuo Yang, Yanyao Shen, Sujay Sanghavi
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 Distributed Data Summarization under Covariate Shift
Kanthi Sarpatwar, Karthikeyan Shanmugam, Venkata Sitaramagiridharganesh Ganapavarapu, Ashish Jagmohan, Roman Vaculin
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 Fenchel Mini-Max Learning
Chenyang Tao, Liqun Chen, Shuyang Dai, Junya Chen, Ke Bai, Dong Wang, Jianfeng Feng, Wenlian Lu, Georgiy Bobashev, Lawrence Carin
Inference, estimation, sampling and likelihood evaluation are four primary goals of probabilistic modeling. Practical considerations often force modeling approaches to make compromises between these objectives. We present a novel probabilistic learning framework, called Fenchel Mini-Max Learning (FML), that accommodates all four desiderata in a flexible and scalable manner. Our derivation is rooted in classical maximum likelihood estimation, and it overcomes a longstanding challenge that prevents unbiased estimation of unnormalized statistical models. By reformulating MLE as a mini-max game, FML enjoys an unbiased training objective that (i) does not explicitly involve the intractable normalizing constant and (ii) is directly amendable to stochastic gradient descent optimization. To demonstrate the utility of the proposed approach, we consider learning unnormalized statistical models, nonparametric density estimation and training generative models, with encouraging empirical results presented.
****
Optimizing Generalized Rate Metrics with Three Players
Harikrishna Narasimhan, Andrew Cotter, Maya Gupta
We present a general framework for solving a large class of learning problems with non-linear functions of classification rates. This includes problems where one wishes to optimize a non-decomposable performance metric such as the F-measure or G-mean, and constrained training problems where the classifier needs to satisfy non-linear rate constraints such as predictive parity fairness, distribution divergences or churn ratios. We extend previous two-player game approaches for constrained optimization to an approach with three players to decouple the classifier rates from the non-linear objective, and seek to find an equilibrium of the game. Our approach generalizes many existing algorithms, and makes possible new algorithms with more flexibility and tighter handling of non-linear rate constraints. We provide convergence guarantees for convex functions of rates, and show how our methodology can be extended to handle sums-of-ratios of rates. Experiments on different fairness tasks confirm the efficacy of our approach.
****
Stability of Graph Scattering Transforms
Fernando Gama, Alejandro Ribeiro, Joan Bruna
Scattering transforms are non-trainable deep convolutional architectures that exploit the multi-scale resolution of a wavelet filter bank to obtain an appropriate representation of data. More importantly, they are proven invariant to translations, and stable to perturbations that are close to translations. This stability property dons the scattering transform with a robustness to small changes in the metric domain of the data. When considering network data, regular convolutions do not hold since the data domain presents an irregular structure given by th

e network topology. In this work, we extend scattering transforms to network data by using multi-resolution graph wavelets, whose computation can be obtained by means of graph convolutions. Furthermore, we prove that the resulting graph scattering transforms are stable to metric perturbations of the underlying network. This renders graph scattering transforms robust to changes on the network topology, making it particularly useful for cases of transfer learning, topology estimation or time-varying graphs.
****
A Geometric Perspective on Optimal Representations for Reinforcement Learning
Marc Bellemare, Will Dabney, Robert Dadashi, Adrien Ali Taiga, Pablo Samuel Castro, Nicolas Le Roux, Dale Schuurmans, Tor Lattimore, Clare Lyle
We propose a new perspective on representation learning in reinforcement learning based on geometric properties of the space of value functions. From there, we provide formal evidence regarding the usefulness of value functions as auxiliary tasks in reinforcement learning. Our formulation considers adapting the representation to minimize the (linear) approximation of the value function of all stationary policies for a given environment. We show that this optimization reduces to making accurate predictions regarding a special class of value functions which we call adversarial value functions (AVFs). We demonstrate that using value functions as auxiliary tasks corresponds to an expected-error relaxation of our formulation, with AVFs a natural candidate, and identify a close relationship with proto-value functions (Mahadevan, 2005). We highlight characteristics of AVFs and their usefulness as auxiliary tasks in a series of experiments on the four-room domain.
****
More Is Less: Learning Efficient Video Representations by Big-Little Network and Depthwise Temporal Aggregation
Quanfu Fan, Chun-Fu (Richard) Chen, Hilde Kuehne, Marco Pistoia, David Cox
Current state-of-the-art models for video action recognition are mostly based on expensive 3D ConvNets. This results in a need for large GPU clusters to train and evaluate such architectures. To address this problem, we present an lightweight and memory-friendly architecture for action recognition that performs on par with or better than current architectures by using only a fraction of resources. The proposed architecture is based on a combination of a deep subnet operating on low-resolution frames with a compact subnet operating on high-resolution frames, allowing for high efficiency and accuracy at the same time. We demonstrate that our approach achieves a reduction by 3~4 times in FLOPs and ~2 times in memory usage compared to the baseline. This enables training deeper models with more input frames under the same computational budget. To further obviate the need for large-scale 3D convolutions, a temporal aggregation module is proposed to model temporal dependencies in a video at very small additional computational costs. Our models achieve strong performance on several action recognition benchmarks including Kinetics, Something-Something and Moments-in-time. The code and models are available at \url{https://github.com/IBM/bLVNet-TAM}.
****
Provably Efficient Q-learning with Function Approximation via Distribution Shift Error Checking Oracle
Simon S. Du, Yuping Luo, Ruosong Wang, Hanrui 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.
****
Learner-aware Teaching: Inverse Reinforcement Learning with Preferences and Constraints
Sebastian Tschiatschek, Ahana Ghosh, Luis Haug, Rati Devidze, Adish Singla
Inverse reinforcement learning (IRL) enables an agent to learn complex behavior by observing demonstrations from a (near-)optimal policy. The typical assumption is that the learner's goal is to match the teacher’s demonstrated behavior. In this paper, we consider the setting where the learner has its own preferences th

at it additionally takes into consideration. These preferences can for example capture behavioral biases, mismatched worldviews, or physical constraints. We study two teaching approaches: learner-agnostic teaching, where the teacher provides demonstrations from an optimal policy ignoring the learner's preferences, and learner-aware teaching, where the teacher accounts for the learner’s preferences. We design learner-aware teaching algorithms and show that significant performance improvements can be achieved over learner-agnostic teaching.
****
PAC-Bayes Un-Expected Bernstein Inequality
Zakaria Mhammedi, Peter Grünwald, Benjamin Guedj
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.
****
Revisiting the Bethe-Hessian: Improved Community Detection in Sparse Heterogeneous Graphs
Lorenzo Dall'Amico, Romain Couillet, Nicolas Tremblay
Spectral clustering is one of the most popular, yet still incompletely understood, methods for community detection on graphs. This article studies spectral clustering based on the Bethe-Hessian matrix $H_r= (r^2-1)I_n+D-rA$ for sparse heterogeneous graphs (following the degree-corrected stochastic block model) in a two-class setting. For a specific value $r=\zeta$, clustering is shown to be insensitive to the degree heterogeneity. We then study the behavior of the informative eigenvector of H_ζ and, as a result, predict the clustering accuracy. The article concludes with an overview of the generalization to more than two classes along with extensive simulations on synthetic and real networks corroborating our findings.
****
Learning Positive Functions with Pseudo Mirror Descent
Yingxiang Yang, Haoxiang Wang, Negar Kiyavash, Niao He
The nonparametric learning of positive-valued functions appears widely in machine learning, especially in the context of estimating intensity functions of point processes. Yet, existing approaches either require computing expensive projections or semidefinite relaxations, or lack convexity and theoretical guarantees after introducing nonlinear link functions. In this paper, we propose a novel algorithm, pseudo mirror descent, that performs efficient estimation of positive functions within a Hilbert space without expensive projections. The algorithm guarantees positivity by performing mirror descent with an appropriately selected Bregman divergence, and a pseudo-gradient is adopted to speed up the gradient evaluation procedure in practice. We analyze both asymptotic and nonasymptotic convergence of the algorithm. Through simulations, we show that pseudo mirror descent outperforms the state-of-the-art benchmarks for learning intensities of Poisson and multivariate Hawkes processes, in terms of both computational efficiency and accuracy.
****
Censored Semi-Bandits: A Framework for Resource Allocation with Censored Feedback
Arun Verma, Manjesh Hanawal, Arun Rajkumar, Raman Sankaran
In this paper, we study Censored Semi-Bandits, a novel variant of the semi-bandits problem. The learner is assumed to have a fixed amount of resources, which it allocates to the arms at each time step. The loss observed from an arm is random and depends on the amount of resources allocated to it. More specifically, the loss equals zero if the allocation for the arm exceeds a constant (but unknown) threshold that can be dependent on the arm. Our goal is to learn a feasible allocation that minimizes the expected loss. The problem is challenging because the loss distribution and threshold value of each arm are unknown. We study this novel setting by establishing its `equivalence' to Multiple-Play Multi-Armed Bandits (MP-MAB) and Combinatorial Semi-Bandits. Exploiting these equivalences, we derive optimal algorithms for our setting using existing algorithms for MP-MAB and Combinatorial Semi-Bandits. Experiments on synthetically generated data validat

e performance guarantees of the proposed algorithms.
****
Defending Against Neural Fake News
Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi
Recent progress in natural language generation has raised dual-use concerns. While applications like summarization and translation are positive, the underlying technology also might enable adversaries to generate neural fake news: targeted propaganda that closely mimics the style of real news.
****
Robust and Communication-Efficient Collaborative Learning
Amirhossein Reisizadeh, Hossein Taheri, Aryan Mokhtari, Hamed Hassani, Ramtin Pedarsani
We consider a decentralized learning problem, where a set of computing nodes aim at solving a non-convex optimization problem collaboratively. It is well-known that decentralized optimization schemes face two major system bottlenecks: stragglers' delay and communication overhead. In this paper, we tackle these bottlenecks by proposing a novel decentralized and gradient-based optimization algorithm named as QuanTimed-DSGD. Our algorithm stands on two main ideas: (i) we impose a deadline on the local gradient computations of each node at each iteration of the algorithm, and (ii) the nodes exchange quantized versions of their local models. The first idea robustifies to straggling nodes and the second alleviates communication efficiency. The key technical contribution of our work is to prove that with non-vanishing noises for quantization and stochastic gradients, the proposed method exactly converges to the global optimal for convex loss functions, and finds a first-order stationary point in non-convex scenarios. Our numerical evaluations of the QuanTimed-DSGD on training benchmark datasets, MNIST and CIFAR-10, demonstrate speedups of up to 3x in run-time, compared to state-of-the-art decentralized optimization methods.
****
A Self Validation Network for Object-Level Human Attention Estimation
Zehua Zhang, Chen Yu, David Crandall
Due to the foveated nature of the human vision system, people can focus their visual attention on a small region of their visual field at a time, which usually contains only a single object. Estimating this object of attention in first-person (egocentric) videos is useful for many human-centered real-world applications such as augmented reality applications and driver assistance systems. A straightforward solution for this problem is to pick the object whose bounding box is hit by the gaze, where eye gaze point estimation is obtained from a traditional eye gaze estimator and object candidates are generated from an off-the-shelf object detector. However, such an approach can fail because it addresses the where and the what problems separately, despite that they are highly related, chicken-and-egg problems. In this paper, we propose a novel unified model that incorporates both spatial and temporal evidence in identifying as well as locating the attended object in firstperson videos. It introduces a novel Self Validation Module that enforces and leverages consistency of the where and the what concepts. We evaluate on two public datasets, demonstrating that Self Validation Module significantly benefits both training and testing and that our model outperforms the state-of-the-art.
****
Learning Robust Global Representations by Penalizing Local Predictive Power
Haohan Wang, Songwei Ge, Zachary Lipton, Eric P. Xing
Despite their renowned in-domain predictive power, convolutional neural networks are known to rely more on high-frequency patterns that humans deem superficial than on low-frequency patterns that agree better with intuitions about what constitutes category membership. This paper proposes a method for training robust convolutional networks by penalizing the predictive power of the local representations learned by earlier layers. Intuitively, our networks are forced to discard predictive signals such as color and texture that can be gleaned from local receptive fields and to rely instead on the global structures of the image. Across a

battery of synthetic and benchmark domain adaptation tasks, our method confers improved generalization out of the domain. Additionally, to evaluate cross-domain transfer, we introduce ImageNet-Sketch, a new dataset consisting of sketch-like images that matches the ImageNet classification validation set in scale and dimension.
****
Average-Case Averages: Private Algorithms for Smooth Sensitivity and Mean Estimation
Mark Bun, Thomas Steinke
The simplest and most widely applied method for guaranteeing differential privacy is to add instance-independent noise to a statistic of interest that is scaled to its global sensitivity. However, global sensitivity is a worst-case notion that is often too conservative for realized dataset instances. We provide methods for scaling noise in an instance-dependent way and demonstrate that they provide greater accuracy under average-case distributional assumptions. Specifically, we consider the basic problem of privately estimating the mean of a real distribution from i.i.d. samples. The standard empirical mean estimator can have arbitrarily-high global sensitivity. We propose the trimmed mean estimator, which interpolates between the mean and the median, as a way of attaining much lower sensitivity on average while losing very little in terms of statistical accuracy. To privately estimate the trimmed mean, we revisit the smooth sensitivity framework of Nissim, Raskhodnikova, and Smith (STOC 2007), which provides a framework for using instance-dependent sensitivity. We propose three new additive noise distributions which provide concentrated differential privacy when scaled to smooth sensitivity. We provide theoretical and experimental evidence showing that our noise distributions compare favorably to others in the literature, in particular, when applied to the mean estimation problem.
****
A Regularized Approach to Sparse Optimal Policy in Reinforcement Learning
Wenhao Yang, Xiang Li, Zhihua Zhang
We propose and study a general framework for regularized Markov decision processes (MDPs) where the goal is to find an optimal policy that maximizes the expected discounted total reward plus a policy regularization term.
The extant entropy-regularized MDPs can be cast into our framework.
Moreover, under our framework, many regularization terms can bring multi-modality and sparsity, which are potentially useful in reinforcement learning.
In particular, we present sufficient and necessary conditions that induce a sparse optimal policy. We also conduct a full mathematical analysis of the proposed regularized MDPs, including the optimality condition, performance error, and sparseness control. We provide a generic method to devise regularization forms and propose off-policy actor critic algorithms in complex environment settings. We empirically analyze the numerical properties of optimal policies and compare the performance of different sparse regularization forms in discrete and continuous environments.
****
Dynamic Ensemble Modeling Approach to Nonstationary Neural Decoding in Brain-Computer Interfaces
Yu Qi, Bin Liu, Yueming Wang, Gang Pan
Brain-computer interfaces (BCIs) have enabled prosthetic device control by decoding motor movements from neural activities. Neural signals recorded from cortex exhibit nonstationary property due to abrupt noises and neuroplastic changes in brain activities during motor control. Current state-of-the-art neural signal decoders such as Kalman filter assume fixed relationship between neural activities and motor movements, thus will fail if this assumption is not satisfied. We propose a dynamic ensemble modeling (DyEnsemble) approach that is capable of adapting to changes in neural signals by employing a proper combination of decoding functions. The DyEnsemble method firstly learns a set of diverse candidate models. Then, it dynamically selects and combines these models online according to Bayesian updating mechanism. Our method can mitigate the effect of noises and cope with different task behaviors by automatic model switching, thus gives more accur

ate predictions. Experiments with neural data demonstrate that the DyEnsemble method outperforms Kalman filters remarkably, and its advantage is more obvious with noisy signals.
****

First-order methods almost always avoid saddle points: The case of vanishing step-sizes
Ioannis Panageas, Georgios Piliouras, 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.
****

Online Markov Decoding: Lower Bounds and Near-Optimal Approximation Algorithms
Vikas Garg, Tamar Pichkhadze
We resolve the fundamental problem of online decoding with general nth order ergodic Markov chain models. Specifically, we provide deterministic and randomized algorithms whose performance is close to that of the optimal offline algorithm even when latency is small. Our algorithms admit efficient implementation via dynamic programs, and readily extend to (adversarial) non-stationary or time-varying settings. We also establish lower bounds for online methods under latency constraints in both deterministic and randomized settings, and show that no online algorithm can perform significantly better than our algorithms. To our knowledge, our work is the first to analyze general Markov chain decoding under hard constraints on latency. We provide strong empirical evidence to illustrate the potential impact of our work in applications such as gene sequencing.
****

Faster Boosting with Smaller Memory
Julaiti Alafate, Yoav S. Freund
State-of-the-art implementations of boosting, such as XGBoost and LightGBM, can process large training sets extremely fast. However, this performance requires that the memory size is sufficient to hold a 2-3 multiple of the training set size. This paper presents an alternative approach to implementing the boosted trees, which achieves a significant speedup over XGBoost and LightGBM, especially when the memory size is small. This is achieved using a combination of three techniques: early stopping, effective sample size, and stratified sampling. Our experiments demonstrate a 10-100 speedup over XGBoost when the training data is too large to fit in memory.
****

Modelling the Dynamics of Multiagent Q-Learning in Repeated Symmetric Games: a Mean Field Theoretic Approach
Shuyue Hu, Chin-wing Leung, Ho-fung Leung
Modelling the dynamics of multi-agent learning has long been an important research topic, but all of the previous works focus on 2-agent settings and mostly use evolutionary game theoretic approaches. In this paper, we study an n-agent setting with n tends to infinity, such that agents learn their policies concurrently over repeated symmetric bimatrix games with some other agents. Using mean field theory, we approximate the effects of other agents on a single agent by an averaged effect. A Fokker-Planck equation that describes the evolution of the probability distribution of Q-values in the agent population is derived. To the best of our knowledge, this is the first time to show the Q-learning dynamics under an n-agent setting can be described by a system of only three equations. We validate our model through comparisons with agent-based simulations on typical symmetric bimatrix games and different initial settings of Q-values.
****

Information-Theoretic Confidence Bounds for Reinforcement Learning
Xiuyuan Lu, Benjamin Van Roy
We integrate information-theoretic concepts into the design and analysis of optimistic algorithms and Thompson sampling. By making a connection between information-theoretic quantities and confidence bounds, we obtain results that relate the per-period performance of the agent with its information gain about the enviro

nment, thus explicitly characterizing the exploration-exploitation tradeoff. The resulting cumulative regret bound depends on the agent's uncertainty over the environment and quantifies the value of prior information. We show applicability of this approach to several environments, including linear bandits, tabular MDPs, and factored MDPs. These examples demonstrate the potential of a general information-theoretic approach for the design and analysis of reinforcement learning algorithms.
****
Bootstrapping Upper Confidence Bound
Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, Guang Cheng
Upper Confidence Bound (UCB) method is arguably the most celebrated one used in online decision making with partial information feedback. Existing techniques for constructing confidence bounds are typically built upon various concentration inequalities, which thus lead to over-exploration. In this paper, we propose a non-parametric and data-dependent UCB algorithm based on the multiplier bootstrap. To improve its finite sample performance, we further incorporate second-order correction into the above construction. In theory, we derive both problem-dependent and problem-independent regret bounds for multi-armed bandits under a much weaker tail assumption than the standard sub-Gaussianity. Numerical results demonstrate significant regret reductions by our method, in comparison with several baselines in a range of multi-armed and linear bandit problems.
****
DETOX: A Redundancy-based Framework for Faster and More Robust Gradient Aggregation
Shashank Rajput, Hongyi Wang, Zachary Charles, Dimitris Papailiopoulos
To improve the resilience of distributed training to worst-case, or Byzantine node failures, several recent methods have replaced gradient averaging with robust aggregation methods. Such techniques can have high computational costs, often quadratic in the number of compute nodes, and only have limited robustness guarantees. Other methods have instead used redundancy to guarantee robustness, but can only tolerate limited numbers of Byzantine failures. In this work, we present DETOX, a Byzantine-resilient distributed training framework that combines algorithmic redundancy with robust aggregation. DETOX operates in two steps, a filtering step that uses limited redundancy to significantly reduce the effect of Byzantine nodes, and a hierarchical aggregation step that can be used in tandem with any state-of-the-art robust aggregation method. We show theoretically that this leads to a substantial increase in robustness, and has a per iteration runtime that can be nearly linear in the number of compute nodes. We provide extensive experiments over real distributed setups across a variety of large-scale machine learning tasks, showing that DETOX leads to orders of magnitude accuracy and speedup improvements over many state-of-the-art Byzantine-resilient approaches.
****
Differentially Private Covariance Estimation
Kareem Amin, Travis Dick, Alex Kulesza, Andres Munoz, Sergei Vassilvitskii
The covariance matrix of a dataset is a fundamental statistic that can be used for calculating optimum regression weights as well as in many other learning and data analysis settings. For datasets containing private user information, we often want to estimate the covariance matrix in a way that preserves differential privacy. While there are known methods for privately computing the covariance matrix, they all have one of two major shortcomings. Some, like the Gaussian mechanism, only guarantee (epsilon, delta)-differential privacy, leaving a non-trivial probability of privacy failure. Others give strong epsilon-differential privacy guarantees, but are impractical, requiring complicated sampling schemes, and tend to perform poorly on real data.
****
Meta-Reinforced Synthetic Data for One-Shot Fine-Grained Visual Recognition
Satoshi Tsutsui, Yanwei Fu, David Crandall
This paper studies the task of one-shot fine-grained recognition, which suffers from the problem of data scarcity of novel fine-grained classes. To alleviate this problem, a off-the-shelf image generator can be applied to synthesize additio

nal images to help one-shot learning. However, such synthesized images may not be helpful in one-shot fine-grained recognition, due to a large domain discrepancy between synthesized and original images. To this end, this paper proposes a meta-learning framework to reinforce the generated images by original images so that these images can facilitate one-shot learning. Specifically, the generic image generator is updated by few training instances of novel classes; and a Meta Image Reinforcing Network (MetaIRNet) is proposed to conduct one-shot fine-grained recognition as well as image reinforcement. The model is trained in an end-to-end manner, and our experiments demonstrate consistent improvement over baseline on one-shot fine-grained image classification benchmarks.
****
PHYRE: A New Benchmark for Physical Reasoning
Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, Ross Girshick
Understanding and reasoning about physics is an important ability of intelligent agents. We develop the PHYRE benchmark for physical reasoning that contains a set of simple classical mechanics puzzles in a 2D physical environment. The benchmark is designed to encourage the development of learning algorithms that are sample-efficient and generalize well across puzzles. We test several modern learning algorithms on PHYRE and find that these algorithms fall short in solving the puzzles efficiently. We expect that PHYRE will encourage the development of novel sample-efficient agents that learn efficient but useful models of physics. For code and to play PHYRE for yourself, please visit https://player.phyre.ai.
****
Facility Location Problem in Differential Privacy Model Revisited
Yunus Esencayi, Marco Gaboardi, Shi Li, Di 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.
****
Provably robust boosted decision stumps and trees against adversarial attacks
Maksym Andriushchenko, Matthias Hein
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-Based Semi-Supervised Learning with Non-ignorable Non-response
Fan Zhou, Tengfei Li, Haibo Zhou, Hongtu Zhu, Ye Jieping
Graph-based semi-supervised learning is a very powerful tool in classification tasks, while in most existing literature the labelled nodes are assumed to be randomly sampled. When the labelling status depends on the unobserved node response, ignoring the missingness can lead to significant estimation bias and handicap the classifiers. This situation is called non-ignorable non-response. To solve the problem, we propose a Graph-based joint model with Non-ignorable Non-response (GNN), followed by a joint inverse weighting estimation procedure incorporated with sampling imputation approach. Our method is proved to outperform some state-of-art models in both regression and classification problems, by simulations and real analysis on the Cora dataset.
****
Latent Ordinary Differential Equations for Irregularly-Sampled Time Series
Yulia Rubanova, Ricky T. Q. Chen, David K. Duvenaud
Time series with non-uniform intervals occur in many applications, and are difficult to model using standard recurrent neural networks (RNNs). We generalize RNNs to have continuous-time hidden dynamics defined by ordinary differential equations (ODEs), a model we call ODE-RNNs. Furthermore, we use ODE-RNNs to replace the recognition network of the recently-proposed Latent ODE model. Both ODE-RNNs and Latent ODEs can naturally handle arbitrary time gaps between observations, and can explicitly model the probability of observation times using Poisson proce

sses. We show experimentally that these ODE-based models outperform their RNN-based counterparts on irregularly-sampled data.
****
On the Correctness and Sample Complexity of Inverse Reinforcement Learning
Abi Komanduru, Jean Honorio
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 New Distribution on the Simplex with Auto-Encoding Applications
Andrew Stirn, Tony Jebara, David Knowles
We construct a new distribution for the simplex using the Kumaraswamy distribution and an ordered stick-breaking process. We explore and develop the theoretical properties of this new distribution and prove that it exhibits symmetry (exchangeability) under the same conditions as the well-known Dirichlet. Like the Dirichlet, the new distribution is adept at capturing sparsity but, unlike the Dirichlet, has an exact and closed form reparameterization--making it well suited for deep variational Bayesian modeling. We demonstrate the distribution's utility in a variety of semi-supervised auto-encoding tasks. In all cases, the resulting models achieve competitive performance commensurate with their simplicity, use of explicit probability models, and abstinence from adversarial training.
****
Model Selection for Contextual Bandits
Dylan J. Foster, Akshay Krishnamurthy, Haipeng Luo
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-The-Loop Optimization: End-To-End Control And Co-Design Of Soft Robots Through Learned Deep Latent Representations
Andrew Spielberg, Allan Zhao, Yuanming Hu, Tao Du, Wojciech Matusik, Daniela Rus
Soft robots have continuum solid bodies that can deform in an infinite number of ways. Controlling soft robots is very challenging as there are no closed form solutions. We present a learning-in-the-loop co-optimization algorithm in which a latent state representation is learned as the robot figures out how to solve the task. Our solution marries hybrid particle-grid-based simulation with deep, variational convolutional autoencoder architectures that can capture salient features of robot dynamics with high efficacy. We demonstrate our dynamics-aware feature learning algorithm on both 2D and 3D soft robots, and show that it is more robust and faster converging than the dynamics-oblivious baseline. We validate the behavior of our algorithm with visualizations of the learned representation.
****
FreeAnchor: Learning to Match Anchors for Visual Object Detection
Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, Qixiang Ye
Modern CNN-based object detectors assign anchors for ground-truth objects under the restriction of object-anchor Intersection-over-Unit (IoU). In this study, we propose a learning-to-match approach to break IoU restriction, allowing objects to match anchors in a flexible manner. Our approach, referred to as FreeAnchor, updates hand-crafted anchor assignment to "free" anchor matching by formulating detector training as a maximum likelihood estimation (MLE) procedure. FreeAnchor targets at learning features which best explain a class of objects in terms of both classification and localization. FreeAnchor is implemented by optimizing detection customized likelihood and can be fused with CNN-based detectors in a plug-and-play manner. Experiments on MS-COCO demonstrate that FreeAnchor consistently outperforms the counterparts with significant margins.
****
Invariance and identifiability issues for word embeddings
Rachel Carrington, Karthik Bharath, Simon Preston

Word embeddings are commonly obtained as optimisers of a criterion function f of a text corpus, but assessed on word-task performance using a different evaluation function g of the test data. We contend that a possible source of disparity in performance on tasks is the incompatibility between classes of transformations that leave f and g invariant. In particular, word embeddings defined by f are not unique; they are defined only up to a class of transformations to which f is invariant, and this class is larger than the class to which g is invariant. One implication of this is that the apparent superiority of one word embedding over another, as measured by word task performance, may largely be a consequence of the arbitrary elements selected from the respective solution sets. We provide a formal treatment of the above identifiability issue, present some numerical examples, and discuss possible resolutions.
****
SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems
Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel Bowman
In the last year, new models and methods for pretraining and transfer learning have driven striking performance improvements across a range of language understanding tasks. The GLUE benchmark, introduced a little over one year ago, offers a single-number metric that summarizes progress on a diverse set of such tasks, but performance on the benchmark has recently surpassed the level of non-expert humans, suggesting limited headroom for further research. In this paper we present SuperGLUE, a new benchmark styled after GLUE with a new set of more difficult language understanding tasks, a software toolkit, and a public leaderboard. SuperGLUE is available at https://super.gluebenchmark.com.
****
PC-Fairness: A Unified Framework for Measuring Causality-based Fairness
Yongkai Wu, Lu Zhang, Xintao Wu, Hanghang Tong
A recent trend of fair machine learning is to define fairness as causality-based notions which concern the causal connection between protected attributes and decisions. However, one common challenge of all causality-based fairness notions is identifiability, i.e., whether they can be uniquely measured from observational data, which is a critical barrier to applying these notions to real-world situations. In this paper, we develop a framework for measuring different causality-based fairness. We propose a unified definition that covers most of previous causality-based fairness notions, namely the path-specific counterfactual fairness (PC fairness). Based on that, we propose a general method in the form of a constrained optimization problem for bounding the path-specific counterfactual fairness under all unidentifiable situations. Experiments on synthetic and real-world datasets show the correctness and effectiveness of our method.
****
Worst-Case Regret Bounds for Exploration via Randomized Value Functions
Daniel Russo
This paper studies a recent proposal to use randomized value functions to drive exploration in reinforcement learning. These randomized value functions are generated by injecting random noise into the training data, making the approach compatible with many popular methods for estimating parameterized value functions. By providing a worst-case regret bound for tabular finite-horizon Markov decision processes, we show that planning with respect to these randomized value functions can induce provably efficient exploration.
****
Glyce: Glyph-vectors for Chinese Character Representations
Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, Jiwei Li
It is intuitive that NLP tasks for logographic languages like Chinese should benefit from the use of the glyph information in those languages. However, due to the lack of rich pictographic evidence in glyphs and the weak generalization ability of standard computer vision models on character data, an effective way to utilize the glyph information remains to be found.

****
Fast and Provable ADMM for Learning with Generative Priors
Fabian Latorre, Armin eftekhari, Volkan Cevher
In this work, we propose a (linearized) Alternating Direction Method-of-Multipliers (ADMM) algorithm for minimizing a convex function subject to a nonconvex constraint. We focus on the special case where such constraint arises from the specification that a variable should lie in the range of a neural network. This is motivated by recent successful applications of Generative Adversarial Networks (GANs) in tasks like compressive sensing, denoising and robustness against adversarial examples. The derived rates for our algorithm are characterized in terms of certain geometric properties of the generator network, which we show hold for feedforward architectures, under mild assumptions. Unlike gradient descent (GD), it can efficiently handle non-smooth objectives as well as exploit efficient partial minimization procedures, thus being faster in many practical scenarios.
****
GRU-ODE-Bayes: Continuous Modeling of Sporadically-Observed Time Series
Edward De Brouwer, Jaak Simm, Adam Arany, Yves Moreau
Modeling real-world multidimensional time series can be particularly challenging when these are sporadically observed (i.e., sampling is irregular both in time and across dimensions)—such as in the case of clinical patient data. To address these challenges, we propose (1) a continuous-time version of the Gated Recurrent Unit, building upon the recent Neural Ordinary Differential Equations (Chen et al., 2018), and (2) a Bayesian update network that processes the sporadic observations. We bring these two ideas together in our GRU-ODE-Bayes method. We then demonstrate that the proposed method encodes a continuity prior for the latent process and that it can exactly represent the Fokker-Planck dynamics of complex processes driven by a multidimensional stochastic differential equation. Additionally, empirical evaluation shows that our method outperforms the state of the art on both synthetic data and real-world data with applications in healthcare and climate forecast. What is more, the continuity prior is shown to be well suited for low number of samples settings.
****
Stochastic Continuous Greedy ++: When Upper and Lower Bounds Match
Amin Karbasi, Hamed Hassani, Aryan Mokhtari, Zebang Shen
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.
****
General E(2)-Equivariant Steerable CNNs
Maurice Weiler, Gabriele Cesa
The big empirical success of group equivariant networks has led in recent years to the sprouting of a great variety of equivariant network architectures. A particular focus has thereby been on rotation and reflection equivariant CNNs for planar images. Here we give a general description of E(2)-equivariant convolutions in the framework of Steerable CNNs. The theory of Steerable CNNs thereby yields constraints on the convolution kernels which depend on group representations describing the transformation laws of feature spaces. We show that these constraints for arbitrary group representations can be reduced to constraints under irreducible representations. A general solution of the kernel space constraint is given for arbitrary representations of the Euclidean group E(2) and its subgroups. We implement a wide range of previously proposed and entirely new equivariant network architectures and extensively compare their performances. E(2)-steerable convolutions are further shown to yield remarkable gains on CIFAR-10, CIFAR-100 and STL-10 when used as drop in replacement for non-equivariant convolutions.
****
On the convergence of single-call stochastic extra-gradient methods
Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, Panayotis Mertikopoulos

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 nonlinear level sets for dimensionality reduction in function approximation

Guannan Zhang, Jiaxin Zhang, Jacob Hinkle

We developed a Nonlinear Level-set Learning (NLL) method for dimensionality reduction in high-dimensional function approximation with small data. This work is motivated by a variety of design tasks in real-world engineering applications, where practitioners would replace their computationally intensive physical models (e.g., high-resolution fluid simulators) with fast-to-evaluate predictive machine learning models, so as to accelerate the engineering design processes. There are two major challenges in constructing such predictive models: (a) high-dimensional inputs (e.g., many independent design parameters) and (b) small training data, generated by running extremely time-consuming simulations. Thus, reducing the input dimension is critical to alleviate the over-fitting issue caused by data insufficiency. Existing methods, including sliced inverse regression and active subspace approaches, reduce the input dimension by learning a linear coordinate transformation; our main contribution is to extend the transformation approach to a nonlinear regime. Specifically, we exploit reversible networks (RevNets) to learn nonlinear level sets of a high-dimensional function and parameterize its level sets in low-dimensional spaces. A new loss function was designed to utilize samples of the target functions' gradient to encourage the transformed function to be sensitive to only a few transformed coordinates. The NLL approach is demonstrated by applying it to three 2D functions and two 20D functions for showing the improved approximation accuracy with the use of nonlinear transformation, as well as to an 8D composite material design problem for optimizing the buckling-resistance performance of composite shells of rocket inter-stages.
****

Regularized Gradient Boosting

Corinna Cortes, Mehryar Mohri, Dmitry Storcheus

Gradient Boosting (\GB) is a popular and very successful ensemble method for binary trees. While various types of regularization of the base predictors are used with this algorithm, the theory that connects such regularizations with generalization guarantees is poorly understood. We fill this gap by deriving data-dependent learning guarantees for \GB\ used with \emph{regularization}, expressed in terms of the Rademacher complexities of the constrained families of base predictors. We introduce a new algorithm, called \rgb\, that directly benefits from these generalization bounds and that, at every boosting round, applies the \emph{Structural Risk Minimization} principle to search for a base predictor with the best empirical fit versus complexity trade-off.
Inspired by \emph{Randomized Coordinate Descent} we provide a scalable implementation of our algorithm, able to search over large families of base predictors. Finally, we provide experimental results, demonstrating that our algorithm achieves significantly better out-of-sample performance on multiple datasets than the standard \GB\ algorithm used with its regularization.
****

Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models

Vincent LE GUEN, Nicolas THOME

This paper addresses the problem of time series forecasting for non-stationary signals and multiple future steps prediction. To handle this challenging task, we
introduce DILATE (DIstortion Loss including shApe and TimE), a new objective function for training deep neural networks. DILATE aims at accurately predicting sudden changes, and explicitly incorporates two terms supporting precise shape and temporal change detection. We introduce a differentiable loss function suitable
for training deep neural nets, and provide a custom back-prop implementation for

speeding up optimization. We also introduce a variant of DILATE, which provides a smooth generalization of temporally-constrained Dynamic TimeWarping (DTW). Experiments carried out on various non-stationary datasets reveal the very good behaviour of DILATE compared to models trained with the standard Mean Squared Error (MSE) loss function, and also to DTW and variants. DILATE is also agnostic to the choice of the model, and we highlight its benefit for training fully conn
ected
networks as well as specialized recurrent architectures, showing its capacity to improve over state-of-the-art trajectory forecasting approaches.
****
General Proximal Incremental Aggregated Gradient Algorithms: Better and Novel Re
sults under General Scheme
Tao Sun, Yuejiao Sun, Dongsheng Li, Qing Liao
The incremental aggregated gradient algorithm is popular in network optimization and machine learning research. However, the current convergence results require the objective function to be strongly convex. And the existing convergence rates are also limited to linear convergence. Due to the mathematical techniques, the stepsize in the algorithm is restricted by the strongly convex constant, which m
ay
make the stepsize be very small (the strongly convex constant may be small).
****
Explaining Landscape Connectivity of Low-cost Solutions for Multilayer Nets
Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge
, Sanjeev Arora
Mode connectivity is a surprising phenomenon in the loss landscape of deep nets. Optima---at least those discovered by gradient-based optimization---turn out to be connected by simple paths on which the loss function is almost constant. Oft
en, these paths can be chosen to be piece-wise linear, with as few as two segmen
ts.
****
Limitations of the empirical Fisher approximation for natural gradient descent
Frederik Kunstner, Philipp Hennig, Lukas Balles
Natural gradient descent, which preconditions a gradient descent update
with the Fisher information matrix of the underlying statistical model,
is a way to capture partial second-order information.
Several highly visible works have advocated an approximation known as the empiri
cal Fisher,
drawing connections between approximate second-order methods and heuristics like Adam.
We dispute this argument by showing that the empirical Fisher---unlike the Fishe
r---does not generally capture second-order information.
We further argue that the conditions under which the empirical Fisher approaches the Fisher (and the Hessian)
are unlikely to be met in practice, and that, even on simple optimization proble
ms,
the pathologies of the empirical Fisher can have undesirable effects.
****
Fast, Provably convergent IRLS Algorithm for p-norm Linear Regression
Deeksha Adil, Richard Peng, Sushant Sachdeva
Linear regression in Lp-norm is a canonical optimization problem that arises in several applications, including sparse recovery, semi-supervised learning, and s
ignal processing. Generic convex optimization algorithms for solving Lp-regressi
on are slow in practice. Iteratively Reweighted Least Squares (IRLS) is an easy to implement family of algorithms for solving these problems that has been studi
ed for over 50 years. However, these algorithms often diverge for p > 3, and sin
ce the work of Osborne (1985), it has been an open problem whether there is an I
RLS algorithm that converges for p > 3. We propose p-IRLS, the first IRLS algori
thm that provably converges geometrically for any p \in [2,\infty). Our algorith
m is simple to implement and is guaranteed to find a high accuracy solution in a sub-linear number of iterations. Our experiments demonstrate that it performs e

ven better than our theoretical bounds, beats the standard Matlab/CVX implementa tion for solving these problems by 10–50x, and is the fastest among available im plementations in the high-accuracy regime.
****
A Model to Search for Synthesizable Molecules
John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin Segler, José Miguel Hernánde z-Lobato
Deep generative models are able to suggest new organic molecules by generating s trings, trees, and graphs representing their structure. While such models allow one to generate molecules with desirable properties, they give no guarantees tha t the molecules can actually be synthesized in practice. We propose a new molecu le generation model, mirroring a more realistic real-world process, where (a) re actants are selected, and (b) combined to form more complex molecules. More spec ifically, our generative model proposes a bag of initial reactants (selected fro m a pool of commercially-available molecules) and uses a reaction model to predi ct how they react together to generate new molecules. We first show that the mod el can generate diverse, valid and unique molecules due to the useful inductive biases of modeling reactions. Furthermore, our model allows chemists to interrog ate not only the properties of the generated molecules but also the feasibility of the synthesis routes. We conclude by using our model to solve retrosynthesis problems, predicting a set of reactants that can produce a target product.
****
Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness
Saeed Mahloujifar, Xiao Zhang, Mohammad Mahmoody, David Evans
Many recent works have shown that adversarial examples that fool classifiers can be found by minimally perturbing a normal input. Recent theoretical results, st arting with Gilmer et al. (2018b), show that if the inputs are drawn from a conc entrated metric probability space, then adversarial examples with small perturba tion are inevitable. A concentrated space has the property that any subset with $\Omega(1)$ (e.g.,1/100) measure, according to the imposed distribution, has small dist ance to almost all (e.g., 99/100) of the points in the space. It is not clear, however, whether these theoretical results apply to actual distributions such as images. This paper presents a method for empirically measuring and bounding t he concentration of a concrete dataset which is proven to converge to the actual concentration. We use it to empirically estimate the intrinsic robustness to an d L2 and Linfinity perturbations of several image classification benchmarks. Cod e for our experiments is available at https://github.com/xiaozhanguva/Measure-Co ncentration.
****
Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queri es
Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, Vicente Ordo nez
This paper explores the task of interactive image retrieval using natural langua ge queries, where a user progressively provides input queries to refine a set of retrieval results. Moreover, our work explores this problem in the context of c omplex image scenes containing multiple objects. We propose Drill-down, an effec tive framework for encoding multiple queries with an efficient compact state rep resentation that significantly extends current methods for single-round image re trieval.
We show that using multiple rounds of natural language queries as input can be s urprisingly effective to find arbitrarily specific images of complex scenes. Fur thermore, we find that existing image datasets with textual captions can provide a surprisingly effective form of weak supervision for this task. We compare our method with existing sequential encoding and embedding networks, demonstrating superior performance on two proposed benchmarks: automatic image retrieval on a simulated scenario that uses region captions as queries, and interactive image r etrieval using real queries from human evaluators.
****
Provably Efficient Q-Learning with Low Switching Cost

Yu Bai, Tengyang Xie, Nan Jiang, Yu-Xiang 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.
****
Fast and Accurate Least-Mean-Squares Solvers
Alaa Maalouf, Ibrahim Jubran, Dan Feldman
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 Agreement Models for Semi-Supervised Learning
Otilia Stretcu, Krishnamurthy Viswanathan, Dana Movshovitz-Attias, Emmanouil Platanios, Sujith Ravi, Andrew Tomkins
Graph-based algorithms are among the most successful paradigms for solving semi-supervised learning tasks. Recent work on graph convolutional networks and neural graph learning methods has successfully combined the expressiveness of neural networks with graph structures. We propose a technique that, when applied to these methods, achieves state-of-the-art results on semi-supervised learning datasets. Traditional graph-based algorithms, such as label propagation, were designed with the underlying assumption that the label of a node can be imputed from that of the neighboring nodes. However, real-world graphs are either noisy or have edges that do not correspond to label agreement. To address this, we propose Graph Agreement Models (GAM), which introduces an auxiliary model that predicts the probability of two nodes sharing the same label as a learned function of their features. The agreement model is used when training a node classification model by encouraging agreement only for the pairs of nodes it deems likely to have the same label, thus guiding its parameters to better local optima. The classification and agreement models are trained jointly in a co-training fashion. Moreover, GAM can also be applied to any semi-supervised classification problem, by inducing a graph whenever one is not provided. We demonstrate that our method achieves a relative improvement of up to 72% for various node classification models, and obtains state-of-the-art results on multiple established datasets.
****
Generalization in Generative Adversarial Networks: A Novel Perspective from Privacy Protection
Bingzhe Wu, Shiwan Zhao, Chaochao Chen, Haoyang Xu, Li Wang, Xiaolu Zhang, Guangyu Sun, Jun Zhou
In this paper, we aim to understand the generalization properties of generative adversarial networks (GANs) from a new perspective of privacy protection. Theoretically, we prove that a differentially private learning algorithm used for training the GAN does not overfit to a certain degree, i.e., the generalization gap can be bounded. Moreover, some recent works, such as the Bayesian GAN, can be re-interpreted based on our theoretical insight from privacy protection. Quantitatively, to evaluate the information leakage of well-trained GAN models, we perform various membership attacks on these models. The results show that previous Lipschitz regularization techniques are effective in not only reducing the generalization gap but also alleviating the information leakage of the training dataset.
****
Large Scale Structure of Neural Network Loss Landscapes
Stanislav Fort, Stanislaw Jastrzebski
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 Similarity Mitigates Test Set Overuse
Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, Benjamin Recht

Excessive reuse of test data has become commonplace in today's machine learning workflows. Popular benchmarks, competitions, industrial scale tuning, among other applications, all involve test data reuse beyond guidance by statistical confidence bounds. Nonetheless, recent replication studies give evidence that popular benchmarks continue to support progress despite years of extensive reuse. We proffer a new explanation for the apparent longevity of test data: Many proposed models are similar in their predictions and we prove that this similarity mitigates overfitting. Specifically, we show empirically that models proposed for the ImageNet ILSVRC benchmark agree in their predictions well beyond what we can conclude from their accuracy levels alone. Likewise, models created by large scale hyperparameter search enjoy high levels of similarity. Motivated by these empirical observations, we give a non-asymptotic generalization bound that takes similarity into account, leading to meaningful confidence bounds in practical settings.
****
Explicit Planning for Efficient Exploration in Reinforcement Learning
Liangpeng Zhang, Ke Tang, Xin Yao
Efficient exploration is crucial to achieving good performance in reinforcement learning. Existing systematic exploration strategies (R-MAX, MBIE, UCRL, etc.), despite being promising theoretically, are essentially greedy strategies that follow some predefined heuristics. When the heuristics do not match the dynamics of Markov decision processes (MDPs) well, an excessive amount of time can be wasted in travelling through already-explored states, lowering the overall efficiency. We argue that explicit planning for exploration can help alleviate such a problem, and propose a Value Iteration for Exploration Cost (VIEC) algorithm which computes the optimal exploration scheme by solving an augmented MDP. We then present a detailed analysis of the exploration behaviour of some popular strategies, showing how these strategies can fail and spend O(n^2 md) or O(n^2 m + nmd) steps to collect sufficient data in some tower-shaped MDPs, while the optimal exploration scheme, which can be obtained by VIEC, only needs O(nmd), where n, m are the numbers of states and actions and d is the data demand. The analysis not only points out the weakness of existing heuristic-based strategies, but also suggests a remarkable potential in explicit planning for exploration.
****
vGraph: A Generative Model for Joint Community Detection and Node Representation Learning
Fan-Yun Sun, Meng Qu, Jordan Hoffmann, Chin-Wei Huang, Jian Tang
This paper focuses on two fundamental tasks of graph analysis: community detection and node representation learning, which capture the global and local structures of graphs respectively. In existing literature, these two tasks are usually independently studied while they are actually highly correlated. We propose a probabilistic generative model called vGraph to learn community membership and node representation collaboratively. Specifically, we assume that each node can be represented as a mixture of communities, and each community is defined as a multinomial distribution over nodes. Both the mixing coefficients and the community distribution are parameterized by the low-dimensional representations of the nodes and communities. We designed an effective variational inference algorithm for the optimization through backpropagation, which regularizes the community membership of neighboring nodes to be similar in the latent space. Experimental results on multiple real-world graphs show that vGraph is very effective in both community detection and node representation learning, outperforming many competitive baselines in both tasks. We show that the framework of vGraph is quite flexible and can be easily extended to detect hierarchical communities.
****
Can Unconditional Language Models Recover Arbitrary Sentences?
Nishant Subramani, Samuel Bowman, Kyunghyun Cho
Neural network-based generative language models like ELMo and BERT can work effectively as general purpose sentence encoders in text classification without further fine-tuning. Is it possible to adapt them in a similar way for use as general-purpose decoders? For this to be possible, it would need to be the case that

for any target sentence of interest, there is some continuous representation tha
t can be passed to the language model to cause it to reproduce that sentence. We
 set aside the difficult problem of designing an encoder that can produce such r
epresentations and, instead, ask directly whether such representations exist at
all. To do this, we introduce a pair of effective, complementary methods for fee
ding representations into pretrained unconditional language models and a corresp
onding set of methods to map sentences into and out of this representation space
, the reparametrized sentence space. We then investigate the conditions under wh
ich a language model can be made to generate a sentence through the identificati
on of a point in such a space and find that it is possible to recover arbitrary
sentences nearly perfectly with language models and representations of moderate
size.
****
A Kernel Loss for Solving the Bellman Equation
Yihao Feng, Lihong Li, Qiang Liu
Value function learning plays a central role in many state-of-the-art reinforcem
ent
learning algorithms. Many popular algorithms like Q-learning do not optimize
any objective function, but are fixed-point iterations of some variants of Bellm
an
operator that are not necessarily a contraction. As a result, they may easily lo
se
convergence guarantees, as can be observed in practice. In this paper, we propos
e a novel loss function, which can be optimized using standard gradient-based me
thods with guaranteed convergence. The key advantage is that its gradient can be
 easily approximated using sampled transitions, avoiding the need for double sam
ples required by prior algorithms like residual gradient. Our approach may be co
mbined with general function classes such as neural networks, using either on- o
r off-policy data, and is shown to work reliably and effectively in several benc
hmarks, including classic problems where standard algorithms are known to diverg
e.
****
Covariate-Powered Empirical Bayes Estimation
Nikolaos Ignatiadis, Stefan Wager
We study methods for simultaneous analysis of many noisy experiments in the pres
ence
of rich covariate information. The goal of the analyst is to optimally estimate
the true effect underlying
each experiment. Both the noisy experimental results and the auxiliary covariate
s are useful for this purpose,
but neither data source on its own captures all the information available to the
 analyst.
In this paper, we propose a flexible plug-in empirical Bayes estimator that synt
hesizes both sources of information and may leverage any black-box predictive mo
del. We show that our approach is within a constant factor of minimax for a simp
le data-generating model.
Furthermore, we establish robust convergence guarantees for our method that hold
 under considerable
generality, and exhibit promising empirical performance on both real and simulat
ed data.
****
Tight Sample Complexity of Learning One-hidden-layer Convolutional Neural Networ
ks
Yuan Cao, Quanquan Gu
We study the sample complexity of learning one-hidden-layer convolutional neural
 networks (CNNs) with non-overlapping filters. We propose a novel algorithm call
ed approximate gradient descent for training CNNs, and show that, with high prob
ability, the proposed algorithm with random initialization grants a linear conve
rgence to the ground-truth parameters up to statistical precision. Compared with
 existing work, our result applies to general non-trivial, monotonic and Lipschi

tz continuous activation functions including ReLU, Leaky ReLU, Sigmod and Softpl
us etc. Moreover, our sample complexity beats existing results in the dependenc
y of the number of hidden nodes and filter size. In fact, our result matches the
 information-theoretic lower bound for learning one-hidden-layer CNNs with linea
r activation functions, suggesting that our sample complexity is tight. Our theo
retical analysis is backed up by numerical experiments.
****
Non-asymptotic Analysis of Stochastic Methods for Non-Smooth Non-Convex Regulari
zed Problems
Yi Xu, Rong Jin, Tianbao Yang
Stochastic Proximal Gradient (SPG) methods have been widely used for solving opt
imization problems with a simple (possibly non-smooth) regularizer in machine le
arning and statistics. However, to the best of our knowledge no non-asymptotic
convergence analysis of SPG exists for non-convex optimization with a non-smooth
 and non-convex regularizer. All existing non-asymptotic analysis of SPG for so
lving non-smooth non-convex problems require the non-smooth regularizer to be a
convex function, and hence are not applicable to a non-smooth non-convex regular
ized problem. This work initiates the analysis to bridge this gap and opens the
door to non-asymptotic convergence analysis of non-smooth non-convex regularized
 problems. We analyze several variants of mini-batch SPG methods for minimizing
a non-convex objective that consists of a smooth non-convex loss and a non-smoot
h non-convex regularizer. Our contributions are two-fold: (i) we show that they
enjoy the same complexities as their counterparts for solving convex regularized
 non-convex problems in terms of finding an approximate stationary point; (ii) w
e develop more practical variants using dynamic mini-batch size instead of a fix
ed mini-batch size without requiring the target accuracy level of solution. Th
e significance of our results is that they improve upon the-state-of-art resul
ts for solving non-smooth non-convex regularized problems. We also empirically d
emonstrate the effectiveness of the considered SPG methods in comparison with ot
her peer stochastic methods.
****
AGEM: Solving Linear Inverse Problems via Deep Priors and Sampling
Bichuan Guo, Yuxing Han, Jiangtao Wen
In this paper we propose to use a denoising autoencoder (DAE) prior to simultane
ously solve a linear inverse problem and estimate its noise parameter. Existing
DAE-based methods estimate the noise parameter empirically or treat it as a tuna
ble hyper-parameter. We instead propose autoencoder guided EM, a probabilistical
ly sound framework that performs Bayesian inference with intractable deep priors
. We show that efficient posterior sampling from the DAE can be achieved via Met
ropolis-Hastings, which allows the Monte Carlo EM algorithm to be used. We demon
strate competitive results for signal denoising, image deblurring and image devi
gnetting. Our method is an example of combining the representation power of deep
 learning with uncertainty quantification from Bayesian statistics.
****
Devign: Effective Vulnerability Identification by Learning Comprehensive Program
 Semantics via Graph Neural Networks
Yaqin Zhou, Shangqing Liu, Jingkai Siow, Xiaoning Du, Yang Liu
Vulnerability identification is crucial to protect the software systems from att
acks
for cyber security. It is especially important to localize the vulnerable functi
ons
among the source code to facilitate the fix. However, it is a challenging and te
dious
process, and also requires specialized security expertise. Inspired by the work
on manually-defined patterns of vulnerabilities from various code representation
graphs and the recent advance on graph neural networks, we propose Devign, a
general graph neural network based model for graph-level classification through
learning on a rich set of code semantic representations. It includes a novel Con
v
module to efficiently extract useful features in the learned rich node represent

ations for graph-level classification. The model is trained over manually labele
d datasets built on 4 diversified large-scale open-source C projects that incorp
orate high complexity and variety of real source code instead of synthesis code
used in previous works. The results of the extensive evaluation on the datasets
demonstrate that Devign outperforms the state of the arts significantly with an
average of 10.51% higher accuracy and 8.68% F1 score, increases averagely 4.66%
accuracy and 6.37% F1 by the Conv module.
****
Probabilistic Watershed: Sampling all spanning forests for seeded segmentation a
nd semi-supervised learning
Enrique Fita Sanmartin, Sebastian Damrich, Fred A. Hamprecht
The seeded Watershed algorithm / minimax semi-supervised learning on a graph com
putes a minimum spanning forest which connects every pixel / unlabeled node to a
 seed / labeled node. We propose instead to consider all possible spanning fores
ts and calculate, for every node, the probability of sampling a forest connecti
ng a certain seed with that node. We dub this approach "Probabilistic Watershed"
. Leo Grady (2006) already noted its equivalence to the Random Walker / Harmonic
 energy minimization. We here give a simpler proof of this equivalence and estab
lish the computational feasibility of the Probabilistic Watershed with Kirchhoff
's matrix tree theorem. Furthermore, we show a new connection between the Random
 Walker probabilities and the triangle inequality of the effective resistance. F
inally, we derive a new and intuitive interpretation of the Power Watershed.
****
Learning Robust Options by Conditional Value at Risk Optimization
Takuya Hiraoka, Takahisa Imagawa, Tatsuya Mori, Takashi Onishi, Yoshimasa Tsuruo
ka
Options are generally learned by using an inaccurate environment model (or simul
ator), which contains uncertain model parameters.
While there are several methods to learn options that are robust against the unc
ertainty of model parameters, these methods only consider either the worst case
or the average (ordinary) case for learning options.
This limited consideration of the cases often produces options that do not work
well in the unconsidered case.
In this paper, we propose a conditional value at risk (CVaR)-based method to lea
rn options that work well in both the average and worst cases.
We extend the CVaR-based policy gradient method proposed by Chow and Ghavamzadeh
 (2014) to deal with robust Markov decision processes and then apply the extende
d method to learning robust options.
We conduct experiments to evaluate our method in multi-joint robot control tasks
 (HopperIceBlock, Half-Cheetah, and Walker2D).
Experimental results show that our method produces options that 1) give better w
orst-case performance than the options learned only to minimize the average-case
 loss, and 2) give better average-case performance than the options learned only
 to minimize the worst-case loss.
****
A Generic Acceleration Framework for Stochastic Composite Optimization
Andrei Kulunchakov, Julien Mairal
In this paper, we introduce various mechanisms to obtain accelerated first-order
 stochastic optimization algorithms when the objective function is convex or str
ongly convex. Specifically, we extend the Catalyst approach originally designed
for deterministic objectives to the stochastic setting. Given an optimization me
thod with mild convergence guarantees for strongly convex problems, the challeng
e is to accelerate convergence to a noise-dominated region, and then achieve con
vergence with an optimal worst-case complexity depending on the noise variance o
f the gradients. A side contribution of our work is also a generic analysis that
 can handle inexact proximal operators, providing new insights about the robustn
ess of stochastic algorithms when the proximal operator cannot be exactly comput
ed.
****
A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Ad

aptation
Runzhe Yang, Xingyuan Sun, Karthik Narasimhan
We introduce a new algorithm for multi-objective reinforcement learning (MORL) with linear preferences, with the goal of enabling few-shot adaptation to new tasks. In MORL, the aim is to learn policies over multiple competing objectives whose relative importance (preferences) is unknown to the agent. While this alleviates dependence on scalar reward design, the expected return of a policy can change significantly with varying preferences, making it challenging to learn a single model to produce optimal policies under different preference conditions. We propose a generalized version of the Bellman equation to learn a single parametric representation for optimal policies over the space of all possible preferences. After an initial learning phase, our agent can execute the optimal policy under any given preference, or automatically infer an underlying preference with very few samples. Experiments across four different domains demonstrate the effectiveness of our approach.
****
Communication trade-offs for Local-SGD with large step size
Aymeric Dieuleveut, Kumar Kshitij Patel
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 modular and programmable architecture search
Renato Negrinho, Matthew Gormley, Geoffrey J. Gordon, Darshan Patil, Nghia Le, Daniel Ferreira
Neural architecture search methods are able to find high performance deep learning architectures with minimal effort from an expert. However, current systems focus on specific use-cases (e.g. convolutional image classifiers and recurrent language models), making them unsuitable for general use-cases that an expert might wish to write. Hyperparameter optimization systems are general-purpose but lack the constructs needed for easy application to architecture search. In this work, we propose a formal language for encoding search spaces over general computational graphs. The language constructs allow us to write modular, composable, and reusable search space encodings and to reason about search space design. We use our language to encode search spaces from the architecture search literature. The language allows us to decouple the implementations of the search space and the search algorithm, allowing us to expose search spaces to search algorithms through a consistent interface. Our experiments show the ease with which we can experiment with different combinations of search spaces and search algorithms without having to implement each combination from scratch. We release an implementation of our language with this paper.
****
Large-scale optimal transport map estimation using projection pursuit
Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, Ping Ma
This paper studies the estimation of large-scale optimal transport maps (OTM), which is a well known challenging problem owing to the curse of dimensionality.
Existing literature approximates the large-scale OTM by a series of one-dimensional OTM problems through iterative random projection.
Such methods, however, suffer from slow or none convergence in practice due to the nature of randomly selected projection directions.
Instead, we propose an estimation method of large-scale OTM by combining the idea of projection pursuit regression and sufficient dimension reduction.
The proposed method, named projection pursuit Monge map (PPMM), adaptively selects the most informative'' projection direction in each iteration.
We theoretically show the proposed dimension reduction method can consistently estimate the mostinformative'' projection direction in each iteration.
Furthermore, the PPMM algorithm weakly convergences to the target large-scale OTM in a reasonable number of steps.
Empirically, PPMM is computationally easy and converges fast.

We assess its finite sample performance through the applications of Wasserstein distance estimation and generative models.
****
Understanding Attention and Generalization in Graph Neural Networks
Boris Knyazev, Graham W. Taylor, Mohamed Amer
We aim to better understand attention over nodes in graph neural networks (GNNs) and identify factors influencing its effectiveness. We particularly focus on the ability of attention GNNs to generalize to larger, more complex or noisy graphs. Motivated by insights from the work on Graph Isomorphism Networks, we design simple graph reasoning tasks that allow us to study attention in a controlled environment. We find that under typical conditions the effect of attention is negligible or even harmful, but under certain conditions it provides an exceptional gain in performance of more than 60% in some of our classification tasks. Satisfying these conditions in practice is challenging and often requires optimal initialization or supervised training of attention. We propose an alternative recipe and train attention in a weakly-supervised fashion that approaches the performance of supervised models, and, compared to unsupervised models, improves results on several synthetic as well as real datasets. Source code and datasets are available at https://github.com/bknyaz/graphattentionpool.
****
Superposition of many models into one
Brian Cheung, Alexander Terekhov, Yubei Chen, Pulkit Agrawal, Bruno Olshausen
We present a method for storing multiple models within a single set of parameters. Models can coexist in superposition and still be retrieved individually. In experiments with neural networks, we show that a surprisingly large number of models can be effectively stored within a single parameter instance. Furthermore, each of these models can undergo thousands of training steps without significantly interfering with other models within the superposition. This approach may be viewed as the online complement of compression: rather than reducing the size of a network after training, we make use of the unrealized capacity of a network during training.
****
A Prior of a Googol Gaussians: a Tensor Ring Induced Prior for Generative Models
Maxim Kuznetsov, Daniil Polykovskiy, Dmitry P. Vetrov, Alex Zhebrak
Generative models produce realistic objects in many domains, including text, image, video, and audio synthesis. Most popular models—Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs)—usually employ a standard Gaussian distribution as a prior. Previous works show that the richer family of prior distributions may help to avoid the mode collapse problem in GANs and to improve the evidence lower bound in VAEs. We propose a new family of prior distributions—Tensor Ring Induced Prior (TRIP)—that packs an exponential number of Gaussians into a high-dimensional lattice with a relatively small number of parameters. We show that these priors improve Fréchet Inception Distance for GANs and Evidence Lower Bound for VAEs. We also study generative models with TRIP in the conditional generation setup with missing conditions. Altogether, we propose a novel plug-and-play framework for generative models that can be utilized in any GAN and VAE-like architectures.
****
Beating SGD Saturation with Tail-Averaging and Minibatching
Nicole Muecke, Gergely Neu, Lorenzo Rosasco
While stochastic gradient descent (SGD) is one of the major workhorses in machine learning, the
learning properties of many practically used variants are still poorly understood.
In this paper, we consider least squares learning in a nonparametric setting and contribute
to filling this gap by focusing on the effect and interplay of multiple passes, mini-batching and
averaging, in particular tail averaging. Our results show how these different variants of SGD

can be combined to achieve optimal learning rates, also providing practical insights. A novel key result is
that tail averaging allows faster convergence rates than uniform averaging in the nonparametric setting.
Further, we show that a combination of tail-averaging and minibatching allows more aggressive
step-size choices than using any one of said components.
****
Extending Stein's unbiased risk estimator to train deep denoisers with correlated pairs of noisy images
Magauiya Zhussip, Shakarim Soltanayev, Se Young Chun
Recently, Stein's unbiased risk estimator (SURE) has been applied to unsupervised training of deep neural network Gaussian denoisers that outperformed classical non-deep learning based denoisers and yielded comparable performance to those trained with ground truth. While SURE requires only one noise realization per image for training, it does not take advantage of having multiple noise realizations per image when they are available (e.g., two uncorrelated noise realizations per image for Noise2Noise). Here, we propose an extended SURE (eSURE) to train deep denoisers with correlated pairs of noise realizations per image and applied it to the case with two uncorrelated realizations per image to achieve better performance than SURE based method and comparable results to Noise2Noise. Then, we further investigated the case with imperfect ground truth (i.e., mild noise in ground truth) that may be obtained considering painstaking, time-consuming, and even expensive processes of collecting ground truth images with multiple noisy images. For the case of generating noisy training data by adding synthetic noise to imperfect ground truth to yield correlated pairs of images, our proposed eSURE based training method outperformed conventional SURE based method as well as Noise2Noise. Code is available at https://github.com/Magauiya/Extended_SURE
****
Preference-Based Batch and Sequential Teaching: Towards a Unified View of Models
Farnam Mansouri, Yuxin Chen, Ara Vartanian, Jerry Zhu, Adish Singla
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.
****
Value Function in Frequency Domain and the Characteristic Value Iteration Algorithm
Amir-massoud Farahmand
This paper considers the problem of estimating the distribution of returns in reinforcement learning (i.e., distributional RL problem). It presents a new representational framework to maintain the uncertainty of returns and provides mathematical tools to compute it.
We show that instead of representing a probability distribution function of returns, one can represent their characteristic function instead, the Fourier transform of their distribution. We call the new representation Characteristic Value Function (CVF), which can be interpreted as the frequency domain representation of the probability distribution of returns.
We show that the CVF satisfies a Bellman-like equation, and its corresponding Bellman operator is contraction with respect to certain metrics.
The contraction property allows us to devise an iterative procedure to compute the CVF, which we call Characteristic Value Iteration (CVI). We analyze CVI and its approximate variant and show how approximation errors affect the quality of computed CVF.
****
Communication-Efficient Distributed Learning via Lazily Aggregated Quantized Gradients
Jun Sun, Tianyi Chen, Georgios Giannakis, Zaiyue Yang
The present paper develops a novel aggregated gradient approach for distributed machine learning that adaptively compresses the gradient communication. The key

idea is to first quantize the computed gradients, and then skip less informative quantized gradient communications by reusing outdated gradients. Quantizing and skipping result in 'lazy' worker-server communications, which justifies the term Lazily Aggregated Quantized gradient that is henceforth abbreviated as LAQ. Our LAQ can provably attain the same linear convergence rate as the gradient descent in the strongly convex case, while effecting major savings in the communication overhead both in transmitted bits as well as in communication rounds. Empirically, experiments with real data corroborate a significant communication reduction compared to existing gradient- and stochastic gradient-based algorithms.
****

Twin Auxilary Classifiers GAN
Mingming Gong, Yanwu Xu, Chunyuan Li, Kun Zhang, Kayhan Batmanghelich
Conditional generative models enjoy significant progress over the past few years. One of the popular conditional models is Auxiliary Classifier GAN (AC-GAN) that generates highly discriminative images by extending the loss function of GAN with an auxiliary classifier. However, the diversity of the generated samples by AC-GAN tends to decrease as the number of classes increases. In this paper, we identify the source of low diversity issue theoretically and propose a practical solution to the problem. We show that the auxiliary classifier in AC-GAN imposes perfect separability, which is disadvantageous when the supports of the class distributions have significant overlap. To address the issue, we propose Twin Auxiliary Classifiers Generative Adversarial Net (TAC-GAN) that adds a new player that interacts with other players (the generator and the discriminator) in GAN. Theoretically, we demonstrate that our TAC-GAN can effectively minimize the divergence between generated and real data distributions. Extensive experimental results show that our TAC-GAN can successfully replicate the true data distributions on simulated data, and significantly improves the diversity of class-conditional image generation on real datasets.
****

Online Prediction of Switching Graph Labelings with Cluster Specialists
Mark Herbster, James Robinson
We address the problem of predicting the labeling of a graph in an online setting when the labeling is changing over time. We present an algorithm based on a specialist approach; we develop the machinery of cluster specialists which probabilistically exploits the cluster structure in the graph. Our algorithm has two variants, one of which surprisingly only requires O(log n) time on any trial t on an n-vertex graph, an exponential speed up over existing methods. We prove switching mistake-bound guarantees for both variants of our algorithm. Furthermore these mistake bounds smoothly vary with the magnitude of the change between successive labelings. We perform experiments on Chicago Divvy Bicycle Sharing data and show that our algorithms significantly outperform an existing algorithm (a kernelized Perceptron) as well as several natural benchmarks.
****

AutoPrune: Automatic Network Pruning by Regularizing Auxiliary Parameters
XIA XIAO, Zigeng Wang, Sanguthevar Rajasekaran
Reducing the model redundancy is an important task to deploy complex deep learning models to resource-limited or time-sensitive devices. Directly regularizing or modifying weight values makes pruning procedure less robust and sensitive to the choice of hyperparameters, and it also requires prior knowledge to tune different hyperparameters for different models. To build a better generalized and easy-to-use pruning method, we propose AutoPrune, which prunes the network through optimizing a set of trainable auxiliary parameters instead of original weights. The instability and noise during training on auxiliary parameters will not directly affect weight values, which makes pruning process more robust to noise and less sensitive to hyperparameters. Moreover, we design gradient update rules for auxiliary parameters to keep them consistent with pruning tasks. Our method can automatically eliminate network redundancy with recoverability, relieving the complicated prior knowledge required to design thresholding functions, and reducing the time for trial and error. We evaluate our method with LeNet and VGG-like on MNIST and CIFAR-10 datasets, and with AlexNet, ResNet and MobileNet on ImageNe

t to establish the scalability of our work. Results show that our model achieves state-of-the-art sparsity, e.g. 7%, 23% FLOPs and 310x, 75x compression ratio for LeNet5 and VGG-like structure without accuracy drop, and 200M and 100M FLOPs for MobileNet V2 with accuracy 73.32% and 66.83% respectively.
****

Understanding the Role of Momentum in Stochastic Gradient Methods
Igor Gitman, Hunter Lang, Pengchuan Zhang, Lin Xiao
The use of momentum in stochastic gradient methods has become a widespread practice in machine learning. Different variants of momentum, including heavy-ball momentum, Nesterov's accelerated gradient (NAG), and quasi-hyperbolic momentum (QHM), have demonstrated success on various tasks. Despite these empirical successes, there is a lack of clear understanding of how the momentum parameters affect convergence and various performance measures of different algorithms. In this paper, we use the general formulation of QHM to give a unified analysis of several popular algorithms, covering their asymptotic convergence conditions, stability regions, and properties of their stationary distributions. In addition, by combining the results on convergence rates and stationary distributions, we obtain sometimes counter-intuitive practical guidelines for setting the learning rate and momentum parameters.
****

DAC: The Double Actor-Critic Architecture for Learning Options
Shangtong Zhang, Shimon Whiteson
We reformulate the option framework as two parallel augmented MDPs. Under this novel formulation, all policy optimization algorithms can be used off the shelf to learn intra-option policies, option termination conditions, and a master policy over options. We apply an actor-critic algorithm on each augmented MDP, yielding the Double Actor-Critic (DAC) architecture. Furthermore, we show that, when state-value functions are used as critics, one critic can be expressed in terms of the other, and hence only one critic is necessary. We conduct an empirical study on challenging robot simulation tasks. In a transfer learning setting, DAC outperforms both its hierarchy-free counterpart and previous gradient-based option learning algorithms.
****

Safe Exploration for Interactive Machine Learning
Matteo Turchetta, Felix Berkenkamp, Andreas Krause
In interactive machine learning (IML), we iteratively make decisions and obtain noisy observations of an unknown function. While IML methods, e.g., Bayesian optimization and active learning, have been successful in applications, on real-world systems they must provably avoid unsafe decisions. To this end, safe IML algorithms must carefully learn about a priori unknown constraints without making unsafe decisions. Existing algorithms for this problem learn about the safety of all decisions to ensure convergence. This is sample-inefficient, as it explores decisions that are not relevant for the original IML objective. In this paper, we introduce a novel framework that renders any existing unsafe IML algorithm safe. Our method works as an add-on that takes suggested decisions as input and exploits regularity assumptions in terms of a Gaussian process prior in order to efficiently learn about their safety. As a result, we only explore the safe set when necessary for the IML problem. We apply our framework to safe Bayesian optimization and to safe exploration in deterministic Markov Decision Processes (MDP), which have been analyzed separately before. Our method outperforms other algorithms empirically.
****

Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthesis Planning
Akihiro Kishimoto, Beat Buesser, Bei Chen, Adi Botea
Search techniques, such as Monte Carlo Tree Search (MCTS) and Proof-Number Search (PNS), are effective in playing and solving games. However, the understanding of their performance in industrial applications is still limited. We investigate MCTS and Depth-First Proof-Number (DFPN) Search, a PNS variant, in the domain of Retrosynthetic Analysis (RA).

We find that DFPN's strengths, that justify its success in games, have limited v
alue in RA, and that an enhanced MCTS variant by Segler et al. significantly out
performs DFPN. We address this disadvantage of DFPN in RA with a novel approach
 to combine DFPN with Heuristic Edge Initialization. Our new search algorithm D
FPN-E outperforms the enhanced MCTS in search time by a factor of 3 on average,
with comparable success rates.
****
Learning from Label Proportions with Generative Adversarial Networks
Jiabin Liu, Bo Wang, Zhiquan Qi, YingJie Tian, Yong Shi
In this paper, we leverage generative adversarial networks (GANs) to derive an e
ffective algorithm LLP-GAN for learning from label proportions (LLP), where only
 the bag-level proportional information in labels is available. Endowed with end
-to-end structure, LLP-GAN performs approximation in the light of an adversarial
 learning mechanism, without imposing restricted assumptions on distribution. Ac
cordingly, we can directly induce the final instance-level classifier upon the d
iscriminator. Under mild assumptions, we give the explicit generative representa
tion and prove the global optimality for LLP-GAN. Additionally, compared with ex
isting methods, our work empowers LLP solver with capable scalability inheriting
 from deep models. Several experiments on benchmark datasets demonstrate vivid a
dvantages of the proposed approach.
****
Sparse High-Dimensional Isotonic Regression
David Gamarnik, Julia Gaudio
We consider the problem of estimating an unknown coordinate-wise monotone functi
on given noisy measurements, known as the isotonic regression problem. Often, on
ly a small subset of the features affects the output. This motivates the sparse
isotonic regression setting, which we consider here. We provide an upper bound o
n the expected VC entropy of the space of sparse coordinate-wise monotone functi
ons, and identify the regime of statistical consistency of our estimator. We als
o propose a linear program to recover the active coordinates, and provide theore
tical recovery guarantees. We close with experiments on cancer classification, a
nd show that our method significantly outperforms several standard methods.
****
Neuropathic Pain Diagnosis Simulator for Causal Discovery Algorithm Evaluation
Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, Cheng Zhang
Discovery of causal relations from observational data is essential for many disc
iplines of science and real-world applications. However, unlike other machine le
arning algorithms, whose development has been greatly fostered by a large amount
 of available benchmark datasets, causal discovery algorithms are notoriously di
fficult to be systematically evaluated because few datasets with known ground-tr
uth causal relations are available. In this work, we handle the problem of evalu
ating causal discovery algorithms by building a flexible simulator in the medica
l setting. We develop a neuropathic pain diagnosis simulator, inspired by the fa
ct that the biological processes of neuropathic pathophysiology are well studied
 with well-understood causal influences. Our simulator exploits the causal graph
 of the neuropathic pain pathology and its parameters in the generator are estim
ated from real-life patient cases. We show that the data generated from our simu
lator have similar statistics as real-world data. As a clear advantage, the simu
lator can produce infinite samples without jeopardizing the privacy of real-worl
d patients. Our simulator provides a natural tool for evaluating various types o
f causal discovery algorithms, including those to deal with practical issues in
causal discovery, such as unknown confounders, selection bias, and missing data.
 Using our simulator, we have evaluated extensively causal discovery algorithms
under various settings.
****
Budgeted Reinforcement Learning in Continuous State Space
Nicolas Carrara, Edouard Leurent, Romain Laroche, Tanguy Urvoy, Odalric-Ambrym M
aillard, Olivier Pietquin
A Budgeted Markov Decision Process (BMDP) is an extension of a Markov Decision P
rocess to critical applications requiring safety constraints. It relies on a not

ion of risk implemented in the shape of an upper bound on a constrains violation signal that -- importantly -- can be modified in real-time. So far, BMDPs could only be solved in the case of finite state spaces with known dynamics. This work extends the state-of-the-art to continuous spaces environments and unknown dynamics. We show that the solution to a BMDP is the fixed point of a novel Budgeted Bellman Optimality operator. This observation allows us to introduce natural extensions of Deep Reinforcement Learning algorithms to address large-scale BMDPs. We validate our approach on two simulated applications: spoken dialogue and autonomous driving.
****
Parameter elimination in particle Gibbs sampling
Anna Wigren, Riccardo Sven Risuleo, Lawrence Murray, Fredrik Lindsten
Bayesian inference in state-space models is challenging due to high-dimensional state trajectories. A viable approach is particle Markov chain Monte Carlo (PMCMC), combining MCMC and sequential Monte Carlo to form ``exact approximations'' to otherwise-intractable MCMC methods. The performance of the approximation is limited to that of the exact method. We focus on particle Gibbs (PG) and particle Gibbs with ancestor sampling (PGAS), improving their performance beyond that of the ideal Gibbs sampler (which they approximate) by marginalizing out one or more parameters. This is possible when the parameter(s) has a conjugate prior relationship with the complete data likelihood. Marginalization yields a non-Markov model for inference, but we show that, in contrast to the general case, the methods still scale linearly in time. While marginalization can be cumbersome to implement, recent advances in probabilistic programming have enabled its automation. We demonstrate how the marginalized methods are viable as efficient inference backends in probabilistic programming, and demonstrate with examples in ecology and epidemiology.
****
Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling
Tengyang Xie, Yifei Ma, Yu-Xiang 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.
****
Understanding Sparse JL for Feature Hashing
Meena Jagadeesan
Feature hashing and other random projection schemes are commonly used to reduce the dimensionality of feature vectors. The goal is to efficiently project a high-dimensional feature vector living in R^n into a much lower-dimensional space R^m, while approximately preserving Euclidean norm. These schemes can be constructed using sparse random projections, for example using a sparse Johnson-Lindenstrauss (JL) transform. A line of work introduced by Weinberger et. al (ICML '09) analyzes the accuracy of sparse JL with sparsity 1 on feature vectors with small linfinity-to-l2 norm ratio. Recently, Freksen, Kamma, and Larsen (NeurIPS '18) closed this line of work by proving a tight tradeoff between linfinity-to-l2 norm ratio and accuracy for sparse JL with sparsity 1. In this paper, we demonstrate the benefits of using sparsity s greater than 1 in sparse JL on feature vectors. Our main result is a tight tradeoff between linfinity-to-l2 norm ratio and accuracy for a general sparsity s, that significantly generalizes the result of Freksen et. al. Our result theoretically demonstrates that sparse JL with s > 1 can have significantly better norm-preservation properties on feature vectors than sparse JL with s = 1; we also empirically demonstrate this finding.
****
Planning in entropy-regularized Markov decision processes and games
Jean-Bastien Grill, Omar Darwiche Domingues, Pierre Menard, Remi Munos, Michal Valko
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 Local Regret for Non-convex Online Forecasting
Sergul Aydore, Tianhao Zhu, Dean P. Foster
We consider online forecasting problems for non-convex machine learning models. Forecasting introduces several challenges such as (i) frequent updates are necessary to deal with concept drift issues since the dynamics of the environment change over time, and (ii) the state of the art models are non-convex models. We address these challenges with a novel regret framework. Standard regret measures commonly do not consider both dynamic environment and non-convex models. We introduce a local regret for non-convex models in a dynamic environment. We present an update rule incurring a cost, according to our proposed local regret, which is sublinear in time T. Our update uses time-smoothed gradients. Using a real-world dataset we show that our time-smoothed approach yields several benefits when compared with state-of-the-art competitors: results are more stable against new data; training is more robust to hyperparameter selection; and our approach is more computationally efficient than the alternatives.
****

NAOMI: Non-Autoregressive Multiresolution Sequence Imputation
Yukai Liu, Rose Yu, Stephan Zheng, Eric Zhan, Yisong Yue
Missing value imputation is a fundamental problem in spatiotemporal modeling, from motion tracking to the dynamics of physical systems. Deep autoregressive models suffer from error propagation which becomes catastrophic for imputing long-range sequences. In this paper, we take a non-autoregressive approach and propose a novel deep generative model: Non-AutOregressive Multiresolution Imputation (NAOMI) to impute long-range sequences given arbitrary missing patterns. NAOMI exploits the multiresolution structure of spatiotemporal data and decodes recursively from coarse to fine-grained resolutions using a divide-and-conquer strategy. We further enhance our model with adversarial training. When evaluated extensively on benchmark datasets from systems of both deterministic and stochastic dynamics. NAOMI demonstrates significant improvement in imputation accuracy (reducing average prediction error by 60% compared to autoregressive counterparts) and generalization for long range sequences.
****

Write, Execute, Assess: Program Synthesis with a REPL
Kevin Ellis, Maxwell Nye, Yewen Pu, Felix Sosa, Josh Tenenbaum, Armando Solar-Lezama
We present a neural program synthesis approach integrating components which write, execute, and assess code to navigate the search space of possible programs. We equip the search process with an interpreter or a read-eval-print-loop (REPL), which immediately executes partially written programs, exposing their semantics. The REPL addresses a basic challenge of program synthesis: tiny changes in syntax can lead to huge changes in semantics. We train a pair of models, a policy that proposes the new piece of code to write, and a value function that assesses the prospects of the code written so-far. At test time we can combine these models with a Sequential Monte Carlo algorithm. We apply our approach to two domains: synthesizing text editing programs and inferring 2D and 3D graphics programs.
****

Conformalized Quantile Regression
Yaniv Romano, Evan Patterson, Emmanuel Candes
Conformal prediction is a technique for constructing prediction intervals that attain valid coverage in finite samples, without making distributional assumptions. Despite this appeal, existing conformal methods can be unnecessarily conservative because they form intervals of constant or weakly varying length across the input space. In this paper we propose a new method that is fully adaptive to heteroscedasticity. It combines conformal prediction with classical quantile regression, inheriting the advantages of both. We establish a theoretical guarantee of valid coverage, supplemented by extensive experiments on popular regression datasets. We compare the efficiency of conformalized quantile regression to other

conformal methods, showing that our method tends to produce shorter intervals.
****

Multiagent Evaluation under Incomplete Information
Mark Rowland, Shayegan Omidshafiei, Karl Tuyls, Julien Perolat, Michal Valko, Georgios Piliouras, 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.
****

SpiderBoost and Momentum: Faster Variance Reduction Algorithms
Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, Vahid Tarokh
SARAH and SPIDER are two recently developed stochastic variance-reduced algorithms, and SPIDER has been shown to achieve a near-optimal first-order oracle complexity in smooth nonconvex optimization. However, SPIDER uses an accuracy-dependent stepsize that slows down the convergence in practice, and cannot handle objective functions that involve nonsmooth regularizers. In this paper, we propose SpiderBoost as an improved scheme, which allows to use a much larger constant-level stepsize while maintaining the same near-optimal oracle complexity, and can be extended with proximal mapping to handle composite optimization (which is nonsmooth and nonconvex) with provable convergence guarantee. In particular, we show that proximal SpiderBoost achieves an oracle complexity of O(min{n^{1/2}\epsilon^{-2},\epsilon^{-3}}) in composite nonconvex optimization, improving the state-of-the-art result by a factor of O(min{n^{1/6},\epsilon^{-1/3}}). We further develop a novel momentum scheme to accelerate SpiderBoost for composite optimization, which achieves the near-optimal oracle complexity in theory and substantial improvement in experiments.
****

Mixtape: Breaking the Softmax Bottleneck Efficiently
Zhilin Yang, Thang Luong, Russ R. Salakhutdinov, Quoc V. Le
The softmax bottleneck has been shown to limit the expressiveness of neural language models. Mixture of Softmaxes (MoS) is an effective approach to address such a theoretical limitation, but are expensive compared to softmax in terms of both memory and time. We propose Mixtape, an output layer that breaks the softmax bottleneck more efficiently with three novel techniques—logit space vector gating, sigmoid tree decomposition, and gate sharing. On four benchmarks including language modeling and machine translation, the Mixtape layer substantially improves the efficiency over the MoS layer by 3.5x to 10.5x while obtaining similar performance. A network equipped with Mixtape is only 20% to 34% slower than a softmax-based network with 10-30K vocabulary sizes, and outperforms softmax in perplexity and translation quality.
****

High-Dimensional Optimization in Adaptive Random Subspaces
Jonathan Lacotte, Mert Pilanci, Marco Pavone
We propose a new randomized optimization method for high-dimensional problems which can be seen as a generalization of coordinate descent to random subspaces. We show that an adaptive sampling strategy for the random subspace significantly outperforms the oblivious sampling method, which is the common choice in the recent literature. The adaptive subspace can be efficiently generated by a correlated random matrix ensemble whose statistics mimic the input data. We prove that the improvement in the relative error of the solution can be tightly characterized in terms of the spectrum of the data matrix, and provide probabilistic upper-bounds. We then illustrate the consequences of our theory with data matrices of different spectral decay. Extensive experimental results show that the proposed approach offers significant speed ups in machine learning problems including logistic regression, kernel classification with random convolution layers and shallow neural networks with rectified linear units. Our analysis is based on convex analysis and Fenchel duality, and establishes connections to sketching and randomized matrix decompositions.
****

Flexible information routing in neural populations through stochastic comodulation
Caroline Haimerl, Cristina Savin, Eero Simoncelli
Humans and animals are capable of flexibly switching between a multitude of tasks, each requiring rapid, sensory-informed decision making. Incoming stimuli are processed by a hierarchy of neural circuits consisting of millions of neurons with diverse feature selectivity. At any given moment, only a small subset of these carry task-relevant information.
In principle, downstream processing stages could identify the relevant neurons through supervised learning, but this would require many example trials. Such extensive learning periods are inconsistent with the observed flexibility of humans or animals, who can adjust to changes in task parameters or structure almost immediately.
Here, we propose a novel solution based on functionally-targeted stochastic modulation. It has been observed that trial-to-trial neural activity is modulated by a shared, low-dimensional, stochastic signal that introduces task-irrelevant noise. Counter-intuitively this noise is preferentially targeted towards task-informative neurons, corrupting the encoded signal. However, we hypothesize that this modulation offers a solution to the identification problem, labeling task-informative neurons so as to facilitate decoding. We simulate an encoding population of spiking neurons whose rates are modulated by a shared stochastic signal, and show that a linear decoder with readout weights approximating neuron-specific modulation strength can achieve near-optimal accuracy. Such a decoder allows fast and flexible task-dependent information routing without relying on hardwired knowledge of the task-informative neurons (as in maximum likelihood) or unrealistically many supervised training trials (as in regression).
****
MarginGAN: Adversarial Training in Semi-Supervised Learning
Jinhao Dong, Tong Lin
A Margin Generative Adversarial Network (MarginGAN) is proposed for semi-supervised learning problems. Like Triple-GAN, the proposed MarginGAN consists of three components---a generator, a discriminator and a classifier, among which two forms of adversarial training arise. The discriminator is trained as usual to distinguish real examples from fake examples produced by the generator. The new feature is that the classifier attempts to increase the margin of real examples and to decrease the margin of fake examples. On the contrary, the purpose of the generator is yielding realistic and large-margin examples in order to fool the discriminator and the classifier simultaneously. Pseudo labels are used for generated and unlabeled examples in training. Our method is motivated by the success of large-margin classifiers and the recent viewpoint that good semi-supervised learning requires a ``bad'' GAN. Experiments on benchmark datasets testify that MarginGAN is orthogonal to several state-of-the-art methods, offering improved error rates and shorter training time as well.
****
Cold Case: The Lost MNIST Digits
Chhavi Yadav, Leon Bottou
Although the popular MNIST dataset \citep{mnist} is derived from the NIST database \citep{nist-sd19}, precise processing steps of this derivation have been lost to time. We propose a reconstruction that is accurate enough to serve as a replacement for the MNIST dataset, with insignificant changes in accuracy. We trace each MNIST digit to its NIST source and its rich metadata such as writer identifier, partition identifier, etc. We also reconstruct the complete MNIST test set with 60,000 samples instead of the usual 10,000. Since the balance 50,000 were never distributed, they enable us to investigate the impact of twenty-five years of MNIST experiments on the reported testing performances. Our results unambiguously confirm the trends observed by \citet{recht2018cifar,recht2019imagenet}: although the misclassification rates are slightly off, classifier ordering and model selection remain broadly reliable. We attribute this phenomenon to the pairing benefits of comparing classifiers on the same digits.
****

RUBi: Reducing Unimodal Biases for Visual Question Answering
Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, Devi Parikh
Visual Question Answering (VQA) is the task of answering questions about an image.
Some VQA models often exploit unimodal biases to provide the correct answer without using the image information.
As a result, they suffer from a huge drop in performance when evaluated on data outside their training set distribution. This critical issue makes them unsuitable for real-world settings.
****
Text-Based Interactive Recommendation via Constraint-Augmented Reinforcement Learning
Ruiyi Zhang, Tong Yu, Yilin Shen, Hongxia Jin, Changyou Chen
Text-based interactive recommendation provides richer user preferences and has demonstrated advantages over traditional interactive recommender systems. However, recommendations can easily violate preferences of users from their past natural-language feedback, since the recommender needs to explore new items for further improvement. To alleviate this issue, we propose a novel constraint-augmented reinforcement learning (RL) framework to efficiently incorporate user preferences over time. Specifically, we leverage a discriminator to detect recommendations violating user historical preference, which is incorporated into the standard RL objective of maximizing expected cumulative future rewards. Our proposed framework is general and is further extended to the task of constrained text generation. Empirical results show that the proposed method yields consistent improvement relative to standard RL methods.
****
Learning to Correlate in Multi-Player General-Sum Sequential Games
Andrea Celli, Alberto Marchesi, Tommaso Bianchi, Nicola Gatti
In the context of multi-player, general-sum games, there is a growing interest in solution concepts involving some form of communication among players, since they can lead to socially better outcomes with respect to Nash equilibria and may be reached through learning dynamics in a decentralized fashion. In this paper, we focus on coarse correlated equilibria (CCEs) in sequential games. First, we complete the picture on the complexity of finding social-welfare-maximizing CCEs by proving that the problem is not in Poly-APX, unless P = NP, in games with three or more players (including chance). Then, we provide simple arguments showing that CFR---working with behavioral strategies---may not converge to a CCE in multi-player, general-sum sequential games. In order to amend this issue, we devise two variants of CFR that provably converge to a CCE. The first one (CFR-S) is a simple stochastic adaptation of CFR which employs sampling to build a correlated strategy, whereas the second variant (called CFR-Jr) enhances CFR with a more involved reconstruction procedure to recover correlated strategies from behavioral ones. Experiments on a rich testbed of multi-player, general-sum sequential games show that both CFR-S and CFR-Jr are dramatically faster than the state-of-the-art algorithms to compute CCEs, with CFR-Jr being also a good heuristic to find socially-optimal CCEs.
****
Learning Sample-Specific Models with Low-Rank Personalized Regression
Ben Lengerich, Bryon Aragam, Eric P. Xing
Modern applications of machine learning (ML) deal with increasingly heterogeneous datasets comprised of data collected from overlapping latent subpopulations. As a result, traditional models trained over large datasets may fail to recognize highly predictive localized effects in favour of weakly predictive global patterns. This is a problem because localized effects are critical to developing individualized policies and treatment plans in applications ranging from precision medicine to advertising. To address this challenge, we propose to estimate sample-specific models that tailor inference and prediction at the individual level. In contrast to classical ML models that estimate a single, complex model (or only a few complex models), our approach produces a model personalized to each sample. These sample-specific models can be studied to understand subgroup dynamics t

hat go beyond coarse-grained class labels. Crucially, our approach does not assume that relationships between samples (e.g. a similarity network) are known a priori. Instead, we use unmodeled covariates to learn a latent distance metric over the samples. We apply this approach to financial, biomedical, and electoral data as well as simulated data and show that sample-specific models provide fine-grained interpretations of complicated phenomena without sacrificing predictive accuracy compared to state-of-the-art models such as deep neural networks.
****

Learning Reward Machines for Partially Observable Reinforcement Learning
Rodrigo Toro Icarte, Ethan Waldie, Toryn Klassen, Rick Valenzano, Margarita Castro, Sheila McIlraith
Reward Machines (RMs), originally proposed for specifying problems in Reinforcement Learning (RL), provide a structured, automata-based representation of a reward function that allows an agent to decompose problems into subproblems that can be efficiently learned using off-policy learning. Here we show that RMs can be learned from experience, instead of being specified by the user, and that the resulting problem decomposition can be used to effectively solve partially observable RL problems. We pose the task of learning RMs as a discrete optimization problem where the objective is to find an RM that decomposes the problem into a set of subproblems such that the combination of their optimal memoryless policies is an optimal policy for the original problem. We show the effectiveness of this approach on three partially observable domains, where it significantly outperforms A3C, PPO, and ACER, and discuss its advantages, limitations, and broader potential.
****

Addressing Sample Complexity in Visual Tasks Using HER and Hallucinatory GANs
Himanshu Sahni, Toby Buckley, Pieter Abbeel, Ilya Kuzovkin
Reinforcement Learning (RL) algorithms typically require millions of environment interactions to learn successful policies in sparse reward settings. Hindsight Experience Replay (HER) was introduced as a technique to increase sample efficiency by reimagining unsuccessful trajectories as successful ones by altering the originally intended goals. However, it cannot be directly applied to visual environments where goal states are often characterized by the presence of distinct visual features. In this work, we show how visual trajectories can be hallucinated to appear successful by altering agent observations using a generative model trained on relatively few snapshots of the goal.
We then use this model in combination with HER to train RL agents in visual settings. We validate our approach on 3D navigation tasks and a simulated robotics application and show marked improvement over baselines derived from previous work.
****

Bat-G net: Bat-inspired High-Resolution 3D Image Reconstruction using Ultrasonic Echoes
Gunpil Hwang, Seohyeon Kim, Hyeon-Min Bae
In this paper, a bat-inspired high-resolution ultrasound 3D imaging system is presented. Live bats demonstrate that the properly used ultrasound can be used to perceive 3D space. With this in mind, a neural network referred to as a Bat-G network is implemented to reconstruct the 3D representation of target objects from the hyperbolic FM (HFM) chirped ultrasonic echoes. The Bat-G network consists of an encoder emulating a bat's central auditory pathway, and a 3D graphical visualization decoder. For the acquisition of the ultrasound data, a custom-made Bat-I sensor module is used. The Bat-G network shows the uniform 3D reconstruction results and achieves precision, recall, and F1-score of 0.896, 0.899 and 0.895, respectively. The experimental results demonstrate the implementation feasibility of a high-resolution non-optical sound-based imaging system being used by live bats. The project web page (https://sites.google.com/view/batgnet) contains additional content summarizing our research.
****

Procrastinating with Confidence: Near-Optimal, Anytime, Adaptive Algorithm Configuration

Robert Kleinberg, Kevin Leyton-Brown, Brendan Lucier, Devon Graham
Algorithm configuration methods optimize the performance of a parameterized heur
istic algorithm on a given distribution of problem instances. Recent work introd
uced an algorithm configuration procedure (Structured Procrastination'') that pr
ovably achieves near optimal performance with high probability and with nearly m
inimal runtime in the worst case. It also offers an anytime property: it keeps t
ightening its optimality guarantees the longer it is run. Unfortunately, Structu
red Procrastination is not adaptive to characteristics of the parameterized algo
rithm: it treats every input like the worst case. Follow-up work (LeapsAndBounds
'') achieves adaptivity but trades away the anytime property. This paper introdu
ces a new algorithm, ``Structured Procrastination with Confidence'', that preser
ves the near-optimality and anytime properties of Structured Procrastination whi
le adding adaptivity. In particular, the new algorithm will perform dramatically
 faster in settings where many algorithm configurations perform poorly. We show
empirically both that such settings arise frequently in practice and that the an
ytime property is useful for finding good configurations quickly.
****
Unsupervised Scalable Representation Learning for Multivariate Time Series
Jean-Yves Franceschi, Aymeric Dieuleveut, Martin Jaggi
Time series constitute a challenging data type for machine learning algorithms,
due to their highly variable lengths and sparse labeling in practice. In this pa
per, we tackle this challenge by proposing an unsupervised method to learn unive
rsal embeddings of time series. Unlike previous works, it is scalable with respe
ct to their length and we demonstrate the quality, transferability and practicab
ility of the learned representations with thorough experiments and comparisons.
To this end, we combine an encoder based on causal dilated convolutions with a n
ovel triplet loss employing time-based negative sampling, obtaining general-purp
ose representations for variable length and multivariate time series.
****
Correlated Uncertainty for Learning Dense Correspondences from Noisy Labels
Natalia Neverova, David Novotny, Andrea Vedaldi
Many machine learning methods depend on human supervision to achieve optimal per
formance. However, in tasks such as DensePose, where the goal is to establish de
nse visual correspondences between images, the quality of manual annotations is
intrinsically limited. We address this issue by augmenting neural network predic
tors with the ability to output a distribution over labels, thus explicitly and
introspectively capturing the aleatoric uncertainty in the annotations.
Compared to previous works, we show that correlated error fields arise naturally
 in applications such as DensePose and these fields can be modeled by deep netwo
rks, leading to a better understanding of the annotation errors.
We show that these models, by understanding uncertainty better, can solve the or
iginal DensePose task more accurately, thus setting the new state-of-the-art acc
uracy in this benchmark.
Finally, we demonstrate the utility of the uncertainty estimates in fusing the p
redictions of produced by multiple models, resulting in a better and more princi
pled approach to model ensembling which can further improve accuracy.
****
Total Least Squares Regression in Input Sparsity Time
Huaian Diao, Zhao Song, David Woodruff, Xin 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Bayesian Learning of Sum-Product Networks
Martin Trapp, Robert Peharz, Hong Ge, Franz Pernkopf, Zoubin Ghahramani
Sum-product networks (SPNs) are flexible density estimators and have received si
gnificant attention due to their attractive inference properties. While paramete
r learning in SPNs is well developed, structure learning leaves something to be
desired: Even though there is a plethora of SPN structure learners, most of them

are somewhat ad-hoc and based on intuition rather than a clear learning principle. In this paper, we introduce a well-principled Bayesian framework for SPN structure learning. First, we decompose the problem into i) laying out a computational graph, and ii) learning the so-called scope function over the graph. The first is rather unproblematic and akin to neural network architecture validation. The second represents the effective structure of the SPN and needs to respect the usual structural constraints in SPN, i.e. completeness and decomposability. While representing and learning the scope function is somewhat involved in general, in this paper, we propose a natural parametrisation for an important and widely used special case of SPNs. These structural parameters are incorporated into a Bayesian model, such that simultaneous structure and parameter learning is cast into monolithic Bayesian posterior inference. In various experiments, our Bayesian SPNs often improve test likelihoods over greedy SPN learners. Further, since the Bayesian framework protects against overfitting, we can evaluate hyper-parameters directly on the Bayesian model score, waiving the need for a separate validation set, which is especially beneficial in low data regimes. Bayesian SPNs can be applied to heterogeneous domains and can easily be extended to nonparametric formulations. Moreover, our Bayesian approach is the first, which consistently and robustly learns SPN structures under missing data.
****
DeepUSPS: Deep Robust Unsupervised Saliency Prediction via Self-supervision
Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, Thomas Brox
Deep neural network (DNN) based salient object detection in images based on high-quality labels is expensive. Alternative unsupervised approaches rely on careful selection of multiple handcrafted saliency methods to generate noisy pseudo-ground-truth labels. In this work, we propose a two-stage mechanism for robust unsupervised object saliency prediction, where the first stage involves refinement of the noisy pseudo labels generated from different handcrafted methods. Each handcrafted method is substituted by a deep network that learns to generate the pseudo labels. These labels are refined incrementally in multiple iterations via our proposed self-supervision technique. In the second stage, the refined labels produced from multiple networks representing multiple saliency methods are used to train the actual saliency detection network. We show that this self-learning procedure outperforms all the existing unsupervised methods over different datasets. Results are even comparable to those of fully-supervised state-of-the-art approaches.
****
Policy Optimization Provably Converges to Nash Equilibria in Zero-Sum Linear Quadratic Games
Kaiqing Zhang, Zhuoran Yang, Tamer Basar
We study the global convergence of policy optimization for finding the Nash equilibria (NE) in zero-sum linear quadratic (LQ) games. To this end, we first investigate the landscape of LQ games, viewing it as a nonconvex-nonconcave saddle-point problem in the policy space. Specifically, we show that despite its nonconvexity and nonconcavity, zero-sum LQ games have the property that the stationary point of the objective function with respect to the linear feedback control policies constitutes the NE of the game. Building upon this, we develop three projected nested-gradient methods that are guaranteed to converge to the NE of the game. Moreover, we show that all these algorithms enjoy both globally sublinear and locally linear convergence rates. Simulation results are also provided to illustrate the satisfactory convergence properties of the algorithms. To the best of our knowledge, this work appears to be the first one to investigate the optimization landscape of LQ games, and provably show the convergence of policy optimization methods to the NE. Our work serves as an initial step toward understanding the theoretical aspects of policy-based reinforcement learning algorithms for zero-sum Markov games in general.
****
On the Power and Limitations of Random Features for Understanding Neural Networks

Gilad Yehudai, Ohad Shamir
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.
****
Real-Time Reinforcement Learning
Simon Ramstedt, Chris Pal
Markov Decision Processes (MDPs), the mathematical framework underlying most algorithms in Reinforcement Learning (RL), are often used in a way that wrongfully assumes that the state of an agent's environment does not change during action selection. As RL systems based on MDPs begin to find application in real-world safety critical situations, this mismatch between the assumptions underlying classical MDPs and the reality of real-time computation may lead to undesirable outcomes.
In this paper, we introduce a new framework, in which states and actions evolve simultaneously and show how it is related to the classical MDP formulation. We analyze existing algorithms under the new real-time formulation and show why they are suboptimal when used in real-time. We then use those insights to create a new algorithm Real-Time Actor Critic (RTAC) that outperforms the existing state-of-the-art continuous control algorithm Soft Actor Critic both in real-time and non-real-time settings.
****
Discriminative Topic Modeling with Logistic LDA
Iryna Korshunova, Hanchen Xiong, Mateusz Fedoryszak, Lucas Theis
Despite many years of research into latent Dirichlet allocation (LDA), applying LDA to collections of non-categorical items is still challenging for practitioners. Yet many problems with much richer data share a similar structure and could benefit from the vast literature on LDA. We propose logistic LDA, a novel discriminative variant of latent Dirichlet allocation which is easy to apply to arbitrary inputs. In particular, our model can easily be applied to groups of images, arbitrary text embeddings, or integrate deep neural networks. Although it is a discriminative model, we show that logistic LDA can learn from unlabeled data in an unsupervised manner by exploiting the group structure present in the data. In contrast to other recent topic models designed to handle arbitrary inputs, our model does not sacrifice the interpretability and principled motivation of LDA.
****
Streaming Bayesian Inference for Crowdsourced Classification
Edoardo Manino, Long Tran-Thanh, Nicholas Jennings
A key challenge in crowdsourcing is inferring the ground truth from noisy and unreliable data. To do so, existing approaches rely on collecting redundant information from the crowd, and aggregating it with some probabilistic method. However, oftentimes such methods are computationally inefficient, are restricted to some specific settings, or lack theoretical guarantees. In this paper, we revisit the problem of binary classification from crowdsourced data. Specifically we propose Streaming Bayesian Inference for Crowdsourcing (SBIC), a new algorithm that does not suffer from any of these limitations. First, SBIC has low complexity and can be used in a real-time online setting. Second, SBIC has the same accuracy as the best state-of-the-art algorithms in all settings. Third, SBIC has provable asymptotic guarantees both in the online and offline settings.
****
Disentangling Influence: Using disentangled representations to audit model predictions
Charles Marx, Richard Phillips, Sorelle Friedler, Carlos Scheidegger, Suresh Venkatasubramanian
Motivated by the need to audit complex and black box models, there has been extensive research on quantifying how data features influence model predictions. Feature influence can be direct (a direct influence on model outcomes) and indirect (model outcomes are influenced via proxy features). Feature influence can also be expressed in aggregate over the training or test data or locally with respect

 to a single point. Current research has typically focused on one of each of these dimensions. In this paper, we develop disentangled influence audits, a procedure to audit the indirect influence of features. Specifically, we show that disentangled representations provide a mechanism to identify proxy features in the dataset, while allowing an explicit computation of feature influence on either individual outcomes or aggregate-level outcomes. We show through both theory and experiments that disentangled influence audits can both detect proxy features and show, for each individual or in aggregate, which of these proxy features affects the classifier being audited the most. In this respect, our method is more powerful than existing methods for ascertaining feature influence.
****
Deep Structured Prediction for Facial Landmark Detection
Lisha Chen, Hui Su, Qiang Ji
Existing deep learning based facial landmark detection methods have achieved excellent performance. These methods, however, do not explicitly embed the structural dependencies among landmark points. They hence cannot preserve the geometric relationships between landmark points or generalize well to challenging conditions or unseen data. This paper proposes a method for deep structured facial landmark detection based on combining a deep Convolutional Network with a Conditional Random Field. We demonstrate its superior performance to existing state-of-the-art techniques in facial landmark detection, especially a better generalization ability on challenging datasets that include large pose and occlusion.
****
Mutually Regressive Point Processes
Ifigeneia Apostolopoulou, Scott Linderman, Kyle Miller, Artur Dubrawski
Many real-world data represent sequences of interdependent events unfolding over time. They can be modeled naturally as realizations of a point process. Despite many potential applications, existing point process models are limited in their ability to capture complex patterns of interaction. Hawkes processes admit many efficient inference algorithms, but are limited to mutually excitatory effects. Non-
linear Hawkes processes allow for more complex influence patterns, but for their estimation it is typically necessary to resort to discrete-time approximations that may yield poor generative models. In this paper, we introduce the first general
class of Bayesian point process models extended with a nonlinear component that allows both excitatory and inhibitory relationships in continuous time. We derive a fully Bayesian inference algorithm for these processes using Polya-Gamma augmentation and Poisson thinning. We evaluate the proposed model on single
and multi-neuronal spike train recordings. Results demonstrate that the proposed model, unlike existing point process models, can generate biologically-plausible spike trains, while still achieving competitive predictive likelihoods.
****
Demystifying Black-box Models with Symbolic Metamodels
Ahmed M. Alaa, Mihaela van der Schaar
Understanding the predictions of a machine learning model can be as crucial as the model's accuracy in many application domains. However, the black-box nature of most highly-accurate (complex) models is a major hindrance to their interpretability. To address this issue, we introduce the symbolic metamodeling framework — a general methodology for interpreting predictions by converting "black-box" models into "white-box" functions that are understandable to human subjects. A symbolic metamodel is a model of a model, i.e., a surrogate model of a trained (machine learning) model expressed through a succinct symbolic expression that comprises familiar mathematical functions and can be subjected to symbolic manipulation. We parameterize symbolic metamodels using Meijer G-functions — a class of complex-valued contour integrals that depend on scalar parameters, and whose solutions reduce to familiar elementary, algebraic, analytic and closed-form functions for different parameter settings. This parameterization enables efficient optimization of metamodels via gradient descent, and allows discovering the functional forms learned by a machine learning model with minimal a priori assumptions.

 We show that symbolic metamodeling provides an all-encompassing framework for m
odel interpretation — all common forms of global and local explanations of a mod
el can be analytically derived from its symbolic metamodel.
****
SHE: A Fast and Accurate Deep Neural Network for Encrypted Data
Qian Lou, Lei Jiang
Homomorphic Encryption (HE) is one of the most promising security solutions to e
merging Machine Learning as a Service (MLaaS). Several Leveled-HE (LHE)-enabled
Convolutional Neural Networks (LHECNNs) are proposed to implement MLaaS to avoid
 the large bootstrapping overhead. However, prior LHECNNs have to pay significan
t computational overhead but achieve only low inference accuracy, due to their p
olynomial approximation activations and poolings. Stacking many polynomial appro
ximation activation layers in a network greatly reduces the inference accuracy,
since the polynomial approximation activation errors lead to a low distortion of
 the output distribution of the next batch normalization layer. So the polynomia
l approximation activations and poolings have become the obstacle to a fast and
accurate LHECNN model.
****
Non-Cooperative Inverse Reinforcement Learning
Xiangyuan Zhang, Kaiqing Zhang, Erik Miehling, Tamer Basar
Making decisions in the presence of a strategic opponent requires one to take in
to account the opponent’s ability to actively mask its intended objective. To de
scribe such strategic situations, we introduce the non-cooperative inverse reinf
orcement learning (N-CIRL) formalism. The N-CIRL formalism consists of two agent
s with completely misaligned objectives, where only one of the agents knows the
true objective function. Formally, we model the N-CIRL formalism as a zero-sum M
arkov game with one-sided incomplete information. Through interacting with the m
ore informed player, the less informed player attempts to both infer and optimiz
e the true objective function. As a result of the one-sided incomplete informati
on, the multi-stage game can be decomposed into a sequence of single- stage game
s expressed by a recursive formula. Solving this recursive formula yields the va
lue of the N-CIRL game and the more informed player’s equilibrium strategy. Anot
her recursive formula, constructed by forming an auxiliary game, termed the dual
 game, yields the less informed player’s strategy. Building upon these two recur
sive formulas, we develop a computationally tractable algorithm to approximately
 solve for the equilibrium strategies. Finally, we demonstrate the benefits of o
ur N-CIRL formalism over the existing multi-agent IRL formalism via extensive nu
merical simulation in a novel cyber security setting.
****
Competitive Gradient Descent
Florian Schaefer, Anima Anandkumar
We introduce a new algorithm for the numerical computation of Nash equilibria of
 competitive two-player games. Our method is a natural generalization of gradien
t descent to the two-player setting where the update is given by the Nash equili
brium of a regularized bilinear local approximation of the underlying game. It
avoids oscillatory and divergent behaviors seen in alternating gradient descent.
 Using numerical experiments and rigorous analysis, we provide a detailed compar
ison to methods based on \emph{optimism} and \emph{consensus} and show that our
 method avoids making any unnecessary changes to the gradient dynamics while ach
ieving exponential (local) convergence for (locally) convex-concave zero sum gam
es. Convergence and stability properties of our method are robust to strong inte
ractions between the players, without adapting the stepsize, which is not the ca
se with previous methods. In our numerical experiments on non-convex-concave pro
blems, existing methods are prone to divergence and instability due to their sen
sitivity to interactions among the players, whereas we never observe divergence
of our algorithm. The ability to choose larger stepsizes furthermore allows our
algorithm to achieve faster convergence, as measured by the number of model eval
uations.
****
Learning in Generalized Linear Contextual Bandits with Stochastic Delays

Zhengyuan Zhou, Renyuan Xu, Jose Blanchet
In this paper, we consider online learning in generalized linear contextual bandits where rewards are not immediately observed. Instead, rewards are available to the decision maker only after some delay, which is unknown and stochastic, even though a decision must be made at each time step for an incoming set of contexts. We study the performance of upper confidence bound (UCB) based algorithms adapted to this delayed setting. In particular, we design a delay-adaptive algorithm, which we call Delayed UCB, for generalized linear contextual bandits using UCB-style exploration and establish regret bounds under various delay assumptions. In the important special case of linear contextual bandits, we further modify this algorithm and establish a tighter regret bound under the same delay assumptions.
Our results contribute to the broad landscape of contextual bandits literature by establishing that UCB algorithms, which are widely deployed in modern recommendation engines, can be made robust to delays.
****
Arbicon-Net: Arbitrary Continuous Geometric Transformation Networks for Image Registration
Jianchun Chen, Lingjing Wang, Xiang Li, Yi Fang
This paper concerns the undetermined problem of estimating geometric transformation between image pairs. Recent methods introduce deep neural networks to predict the controlling parameters of hand-crafted geometric transformation models (e.g. thin-plate spline) for image registration and matching. However, the low-dimension parametric models are incapable of estimating a highly complex geometric transform with limited flexibility to model the actual geometric deformation from image pairs. To address this issue, we present an end-to-end trainable deep neural networks, named Arbitrary Continuous Geometric Transformation Networks (Arbicon-Net), to directly predict the dense displacement field for pairwise image alignment. Arbicon-Net is generalized from training data to predict the desired arbitrary continuous geometric transformation in a data-driven manner for unseen new pair of images. Particularly, without imposing penalization terms, the predicted displacement vector function is proven to be spatially continuous and smooth. To verify the performance of Arbicon-Net, we conducted semantic alignment tests over both synthetic and real image dataset with various experimental settings. The results demonstrate that Arbicon-Net outperforms the previous image alignment techniques in identifying the image correspondences.
****
On the Calibration of Multiclass Classification with Rejection
Chenri Ni, Nontawat Charoenphakdee, Junya Honda, Masashi Sugiyama
We investigate the problem of multiclass classification with rejection, where a classifier can choose not to make a prediction to avoid critical misclassification. First, we consider an approach based on simultaneous training of a classifier and a rejector, which achieves the state-of-the-art performance in the binary case. We analyze this approach for the multiclass case and derive a general condition for calibration to the Bayes-optimal solution, which suggests that calibration is hard to achieve by general loss functions unlike the binary case. Next, we consider another traditional approach based on confidence scores, in which the existing work focuses on a specific class of losses. We propose rejection criteria for more general losses for this approach and guarantee calibration to the Bayes-optimal solution. Finally, we conduct experiments to validate the relevance of our theoretical findings.
****
Point-Voxel CNN for Efficient 3D Deep Learning
Zhijian Liu, Haotian Tang, Yujun Lin, Song Han
We present Point-Voxel CNN (PVCNN) for efficient, fast 3D deep learning. Previous work processes 3D data using either voxel-based or point-based NN models. However, both approaches are computationally inefficient. The computation cost and memory footprints of the voxel-based models grow cubically with the input resolution, making it memory-prohibitive to scale up the resolution. As for point-based networks, up to 80% of the time is wasted on dealing with the sparse data which

have rather poor memory locality, not on the actual feature extraction. In this paper, we propose PVCNN that represents the 3D input data in points to reduce the memory consumption, while performing the convolutions in voxels to reduce the irregular, sparse data access and improve the locality. Our PVCNN model is both memory and computation efficient. Evaluated on semantic and part segmentation datasets, it achieves much higher accuracy than the voxel-based baseline with 10× GPU memory reduction; it also outperforms the state-of-the-art point-based models with 7× measured speedup on average. Remarkably, the narrower version of PVCNN achieves 2× speedup over PointNet (an extremely efficient model) on part and scene segmentation benchmarks with much higher accuracy. We validate the general effectiveness of PVCNN on 3D object detection: by replacing the primitives in Frustrum PointNet with PVConv, it outperforms Frustrum PointNet++ by 2.4% mAP on average with 1.5× measured speedup and GPU memory reduction.
****
Importance Weighted Hierarchical Variational Inference
Artem Sobolev, Dmitry P. Vetrov
Variational Inference is a powerful tool in the Bayesian modeling toolkit, however, its effectiveness is determined by the expressivity of the utilized variational distributions in terms of their ability to match the true posterior distribution. In turn, the expressivity of the variational family is largely limited by the requirement of having a tractable density function.
To overcome this roadblock, we introduce a new family of variational upper bounds on a marginal log-density in the case of hierarchical models (also known as latent variable models). We then derive a family of increasingly tighter variational lower bounds on the otherwise intractable standard evidence lower bound for hierarchical variational distributions, enabling the use of more expressive approximate posteriors. We show that previously known methods, such as Hierarchical Variational Models, Semi-Implicit Variational Inference and Doubly Semi-Implicit Variational Inference can be seen as special cases of the proposed approach, and empirically demonstrate superior performance of the proposed method in a set of experiments.
****
Fast Convergence of Belief Propagation to Global Optima: Beyond Correlation Decay
Frederic Koehler
Belief propagation is a fundamental message-passing algorithm for probabilistic reasoning and inference in graphical models. While it is known to be exact on trees, in most applications belief propagation is run on graphs with cycles. Understanding the behavior of loopy'' belief propagation has been a major challenge for researchers in machine learning, and several positive convergence results for BP are known under strong assumptions which imply the underlying graphical model exhibits decay of correlations. We show that under a natural initialization, BP converges quickly to the global optimum of the Bethe free energy for Ising models on arbitrary graphs, as long as the Ising model is \emph{ferromagnetic} (i.e. neighbors prefer to be aligned). This holds even though such models can exhibit long range correlations and may have multiple suboptimal BP fixed points. We also show an analogous result for iterating the (naive) mean-field equations; perhaps surprisingly, both results aredimension-free'' in the sense that a constant number of iterations already provides a good estimate to the Bethe/mean-field free energy.
****
ZO-AdaMM: Zeroth-Order Adaptive Momentum Method for Black-Box Optimization
Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, David Cox
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.
****
U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging

Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, Christian Ige
l
Neural networks are becoming more and more popular for the analysis of physiolog
ical time-series. The most successful deep learning systems in this domain combi
ne convolutional and recurrent layers to extract useful features to model tempor
al relations. Unfortunately, these recurrent models are difficult to tune and op
timize. In our experience, they often require task-specific modifications, which
 makes them challenging to use for non-experts. We propose U-Time, a fully feed-
forward deep learning approach to physiological time series segmentation develop
ed for the analysis of sleep data. U-Time is a temporal fully convolutional netw
ork based on the U-Net architecture that was originally proposed for image segme
ntation. U-Time maps sequential inputs of arbitrary length to sequences of class
 labels on a freely chosen temporal scale. This is done by implicitly classifyin
g every individual time-point of the input signal and aggregating these classifi
cations over fixed intervals to form the final predictions. We evaluated U-Time
for sleep stage classification on a large collection of sleep electroencephalogr
aphy (EEG) datasets. In all cases, we found that U-Time reaches or outperforms c
urrent state-of-the-art deep learning models while being much more robust in the
 training process and without requiring architecture or hyperparameter adaptatio
n across tasks.
****
Meta-Curvature
Eunbyung Park, Junier B. Oliva
We propose meta-curvature (MC), a framework to learn curvature information for b
etter generalization and fast model adaptation. MC expands on the model-agnostic
 meta-learner (MAML) by learning to transform the gradients in the inner optimiz
ation such that the transformed gradients achieve better generalization performa
nce to a new task. For training large scale neural networks, we decompose the cu
rvature matrix into smaller matrices in a novel scheme where we capture the depe
ndencies of the model's parameters with a series of tensor products. We demonstr
ate the effects of our proposed method on several few-shot learning tasks and da
tasets. Without any task specific techniques and architectures, the proposed met
hod achieves substantial improvement upon previous MAML variants and outperforms
 the recent state-of-the-art methods. Furthermore, we observe faster convergence
 rates of the meta-training process. Finally, we present an analysis that explai
ns better generalization performance with the meta-trained curvature.
****
Exploration via Hindsight Goal Generation
Zhizhou Ren, Kefan Dong, Yuan Zhou, Qiang Liu, Jian Peng
Goal-oriented reinforcement learning has recently been a practical framework for
 robotic manipulation tasks, in which an agent is required to reach a certain go
al defined by a function on the state space. However, the sparsity of such rewar
d definition makes traditional reinforcement learning algorithms very inefficien
t. Hindsight Experience Replay (HER), a recent advance, has greatly improved sam
ple efficiency and practical applicability for such problems. It exploits previo
us replays by constructing imaginary goals in a simple heuristic way, acting lik
e an implicit curriculum to alleviate the challenge of sparse reward signal. In
this paper, we introduce Hindsight Goal Generation (HGG), a novel algorithmic fr
amework that generates valuable hindsight goals which are easy for an agent to a
chieve in the short term and are also potential for guiding the agent to reach t
he actual goal in the long term. We have extensively evaluated our goal generati
on algorithm on a number of robotic manipulation tasks and demonstrated substant
ially improvement over the original HER in terms of sample efficiency.
****
VIREL: A Variational Inference Framework for Reinforcement Learning
Matthew Fellows, Anuj Mahajan, Tim G. J. Rudner, Shimon Whiteson
Applying probabilistic models to reinforcement learning (RL) enables the uses of
 powerful optimisation tools such as variational inference in RL. However, exist
ing inference frameworks and their algorithms pose significant challenges for le
arning optimal policies, e.g., the lack of mode capturing behaviour in pseudo-li

kelihood methods, difficulties learning deterministic policies in maximum entropy RL based approaches, and a lack of analysis when function approximators are used. We propose VIREL, a theoretically grounded probabilistic inference framework for RL that utilises a parametrised action-value function to summarise future dynamics of the underlying MDP, generalising existing approaches. VIREL also benefits from a mode-seeking form of KL divergence, the ability to learn deterministic optimal polices naturally from inference, and the ability to optimise value functions and policies in separate, iterative steps. In applying variational expectation-maximisation to VIREL, we thus show that the actor-critic algorithm can be reduced to expectation-maximisation, with policy improvement equivalent to an E-step and policy evaluation to an M-step. We then derive a family of actor-critic methods fromVIREL, including a scheme for adaptive exploration. Finally, we demonstrate that actor-critic algorithms from this family outperform state-of-the-art methods based on soft value functions in several domains.

****

What Can ResNet Learn Efficiently, Going Beyond Kernels?

Zeyuan Allen-Zhu, Yuanzhi 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.

****

Trajectory of Alternating Direction Method of Multipliers and Adaptive Acceleration

Clarice Poon, Jingwei Liang

The alternating direction method of multipliers (ADMM) is one of the most widely used first-order optimisation methods in the literature owing to its simplicity, flexibility and efficiency. Over the years, numerous efforts are made to improve the performance of the method, such as the inertial technique. By studying the geometric properties of ADMM, we discuss the limitations of current inertial accelerated ADMM and then present and analyze an adaptive acceleration scheme for the method. Numerical experiments on problems arising from image processing, statistics and machine learning demonstrate the advantages of the proposed acceleration approach.

****

Reducing Noise in GAN Training with Variance Reduced Extragradient

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, Simon Lacoste-Julien

We study the effect of the stochastic gradient noise on the training of generative adversarial networks (GANs) and show that it can prevent the convergence of standard game optimization methods, while the batch version converges. We address this issue with a novel stochastic variance-reduced extragradient (SVRE) optimization algorithm, which for a large class of games improves upon the previous convergence rates proposed in the literature. We observe empirically that SVRE performs similarly to a batch method on MNIST while being computationally cheaper, and that SVRE yields more stable GAN training on standard datasets.

****

Focused Quantization for Sparse CNNs

Yiren Zhao, Xitong Gao, Daniel Bates, Robert Mullins, Cheng-Zhong Xu

Deep convolutional neural networks (CNNs) are powerful tools for a wide range of vision tasks, but the enormous amount of memory and compute resources required by CNNs poses a challenge in deploying them on constrained devices. Existing compression techniques, while excelling at reducing model sizes, struggle to be computationally friendly. In this paper, we attend to the statistical properties of sparse CNNs and present focused quantization, a novel quantization strategy based on power-of-two values, which exploits the weight distributions after fine-grained pruning. The proposed method dynamically discovers the most effective numerical representation for weights in layers with varying sparsities, significantly reducing model sizes. Multiplications in quantized CNNs are replaced with much cheaper bit-shift operations for efficient inference. Coupled with lossless encoding, we build a compression pipeline that provides CNNs with high compression

ratios (CR), low computation cost and minimal loss in accuracies. In ResNet-50, we achieved a 18.08x CR with only 0.24% loss in top-5 accuracy, outperforming existing compression methods. We fully compress a ResNet-18 and found that it is not only higher in CR and top-5 accuracy, but also more hardware efficient as it requires fewer logic gates to implement when compared to other state-of-the-art quantization methods assuming the same throughput.
****
Submodular Function Minimization with Noisy Evaluation Oracle
Shinji Ito
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.
****
Knowledge Extraction with No Observable Data
Jaemin Yoo, Minyong Cho, Taebum Kim, U Kang
Knowledge distillation is to transfer the knowledge of a large neural network into a smaller one and has been shown to be effective especially when the amount of training data is limited or the size of the student model is very small. To transfer the knowledge, it is essential to observe the data that have been used to train the network since its knowledge is concentrated on a narrow manifold rather than the whole input space. However, the data are not accessible in many cases due to the privacy or confidentiality issues in medical, industrial, and military domains. To the best of our knowledge, there has been no approach that distills the knowledge of a neural network when no data are observable. In this work, we propose KegNet (Knowledge Extraction with Generative Networks), a novel approach to extract the knowledge of a trained deep neural network and to generate artificial data points that replace the missing training data in knowledge distillation. Experiments show that KegNet outperforms all baselines for data-free knowledge distillation. We provide the source code of our paper in https://github.com/snudatalab/KegNet.
****
Global Guarantees for Blind Demodulation with Generative Priors
Paul Hand, Babhru Joshi
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 Jump Stochastic Differential Equations
Junteng Jia, Austin R. Benson
Many time series are effectively generated by a combination of deterministic continuous flows along with discrete jumps sparked by stochastic events. However, we usually do not have the equation of motion describing the flows, or how they are affected by jumps. To this end, we introduce Neural Jump Stochastic Differential Equations that provide a data-driven approach to learn continuous and discrete dynamic behavior, i.e., hybrid systems that both flow and jump. Our approach extends the framework of Neural Ordinary Differential Equations with a stochastic process term that models discrete events. We then model temporal point processes with a piecewise-continuous latent trajectory, where the discontinuities are caused by stochastic events whose conditional intensity depends on the latent state. We demonstrate the predictive capabilities of our model on a range of synthetic and real-world marked point process datasets, including classical point processes (such as Hawkes processes), awards on Stack Overflow, medical records, and earthquake monitoring.
****
Intrinsically Efficient, Stable, and Bounded Off-Policy Evaluation for Reinforcement Learning
Nathan Kallus, Masatoshi Uehara
Off-policy evaluation (OPE) in both contextual bandits and reinforcement learnin

g allows one to evaluate novel decision policies without needing to conduct exploration, which is often costly or otherwise infeasible. The problem's importance has attracted many proposed solutions, including importance sampling (IS), self-normalized IS (SNIS), and doubly robust (DR) estimates. DR and its variants ensure semiparametric local efficiency if Q-functions are well-specified, but if they are not they can be worse than both IS and SNIS. It also does not enjoy SNIS's inherent stability and boundedness. We propose new estimators for OPE based on empirical likelihood that are always more efficient than IS, SNIS, and DR and satisfy the same stability and boundedness properties as SNIS. On the way, we categorize various properties and classify existing estimators by them. Besides the theoretical guarantees, empirical studies suggest the new estimators provide advantages.
****
Learning about an exponential amount of conditional distributions
Mohamed Belghazi, Maxime Oquab, David Lopez-Paz
We introduce the Neural Conditioner (NC), a self-supervised machine able to learn about all the conditional distributions of a random vector X. The NC is a function NC(x⋅a,a,r) that leverages adversarial training to match each conditional distribution P(Xr|Xa=xa). After training, the NC generalizes to sample from conditional distributions never seen, including the joint distribution. The NC is also able to auto-encode examples, providing data representations useful for downstream classification tasks. In sum, the NC integrates different self-supervised tasks (each being the estimation of a conditional distribution) and levels of supervision (partially observed data) seamlessly into a single learning experience.
****
Multi-mapping Image-to-Image Translation via Learning Disentanglement
Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, Ge Li
Recent advances of image-to-image translation focus on learning the one-to-many mapping from two aspects: multi-modal translation and multi-domain translation. However, the existing methods only consider one of the two perspectives, which makes them unable to solve each other's problem. To address this issue, we propose a novel unified model, which bridges these two objectives. First, we disentangle the input images into the latent representations by an encoder-decoder architecture with a conditional adversarial training in the feature space. Then, we encourage the generator to learn multi-mappings by a random cross-domain translation. As a result, we can manipulate different parts of the latent representations to perform multi-modal and multi-domain translations simultaneously.
Experiments demonstrate that our method outperforms state-of-the-art methods.
****
Computational Mirrors: Blind Inverse Light Transport by Deep Matrix Factorization
Miika Aittala, Prafull Sharma, Lukas Murmann, Adam Yedidia, Gregory Wornell, Bill Freeman, Fredo Durand
We recover a video of the motion taking place in a hidden scene by observing changes in indirect illumination in a nearby uncalibrated visible region. We solve this problem by factoring the observed video into a matrix product between the unknown hidden scene video and an unknown light transport matrix. This task is extremely ill-posed, as any non-negative factorization will satisfy the data. Inspired by recent work on the Deep Image Prior, we parameterize the factor matrices using randomly initialized convolutional neural networks trained in a one-off manner, and show that this results in decompositions that reflect the true motion in the hidden scene.
****
Explicitly disentangling image content from translation and rotation with spatial-VAE
Tristan Bepler, Ellen Zhong, Kotaro Kelley, Edward Brignole, Bonnie Berger
Given an image dataset, we are often interested in finding data generative factors that encode semantic content independently from pose variables such as rotation and translation. However, current disentanglement approaches do not impose any specific structure on the learned latent representations. We propose a method

for explicitly disentangling image rotation and translation from other unstructured latent factors in a variational autoencoder (VAE) framework. By formulating the generative model as a function of the spatial coordinate, we make the reconstruction error differentiable with respect to latent translation and rotation parameters. This formulation allows us to train a neural network to perform approximate inference on these latent variables while explicitly constraining them to only represent rotation and translation. We demonstrate that this framework, termed spatial-VAE, effectively learns latent representations that disentangle image rotation and translation from content and improves reconstruction over standard VAEs on several benchmark datasets, including applications to modeling continuous 2-D views of proteins from single particle electron microscopy and galaxies in astronomical images.
****
Imitation-Projected Programmatic Reinforcement Learning
Abhinav Verma, Hoang Le, Yisong Yue, Swarat Chaudhuri
We study the problem of programmatic reinforcement learning, in which policies are represented as short programs in a symbolic language. Programmatic policies can be more interpretable, generalizable, and amenable to formal verification than neural policies; however, designing rigorous learning approaches for such policies remains a challenge. Our approach to this challenge - a meta-algorithm called PROPEL - is based on three insights. First, we view our learning task as optimization in policy space, modulo the constraint that the desired policy has a programmatic representation, and solve this optimization problem using a form of mirror descent that takes a gradient step into the unconstrained policy space and then projects back onto the constrained space. Second, we view the unconstrained policy space as mixing neural and programmatic representations, which enables employing state-of-the-art deep policy gradient approaches. Third, we cast the projection step as program synthesis via imitation learning, and exploit contemporary combinatorial methods for this task. We present theoretical convergence results for PROPEL and empirically evaluate the approach in three continuous control domains. The experiments show that PROPEL can significantly outperform state-of-the-art approaches for learning programmatic policies.
****
The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies
Basri Ronen, David Jacobs, Yoni Kasten, Shira Kritchman
We study the relationship between the frequency of a function and the speed at which a neural network learns it. We build on recent results that show that the dynamics of overparameterized neural networks trained with gradient descent can be well approximated by a linear system. When normalized training data is uniformly distributed on a hypersphere, the eigenfunctions of this linear system are spherical harmonic functions. We derive the corresponding eigenvalues for each frequency after introducing a bias term in the model. This bias term had been omitted from the linear network model without significantly affecting previous theoretical results. However, we show theoretically and experimentally that a shallow neural network without bias cannot represent or learn simple, low frequency functions with odd frequencies. Our results lead to specific predictions of the time it will take a network to learn functions of varying frequency. These predictions match the empirical behavior of both shallow and deep networks.
****
Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem
Gonzalo Mena, Jonathan Niles-Weed
We prove several fundamental statistical bounds for entropic OT with the squared Euclidean cost between subgaussian probability measures in arbitrary dimension. First, through a new sample complexity result we establish the rate of convergence of entropic OT for empirical measures.
Our analysis improves exponentially on the bound of Genevay et al.~(2019) and extends their work to unbounded measures.
Second, we establish a central limit theorem for entropic OT, based on technique

s developed by Del Barrio and Loubes~(2019).
Previously, such a result was only known for finite metric spaces.
As an application of our results, we develop and analyze a new technique for est
imating the entropy of a random variable corrupted by gaussian noise.
****
A Game Theoretic Approach to Class-wise Selective Rationalization
Shiyu Chang, Yang Zhang, Mo Yu, Tommi Jaakkola
Selection of input features such as relevant pieces of text has become a common
technique of highlighting how complex neural predictors operate. The selection c
an be optimized post-hoc for trained models or incorporated directly into the me
thod itself (self-explaining). However, an overall selection does not properly c
apture the multi-faceted nature of useful rationales such as pros and cons for d
ecisions. To this end, we propose a new game theoretic approach to class-depende
nt rationalization, where the method is specifically trained to highlight eviden
ce supporting alternative conclusions. Each class involves three players set up
competitively to find evidence for factual and counterfactual scenarios. We show
 theoretically in a simplified scenario how the game drives the solution towards
 meaningful class-dependent rationales. We evaluate the method in single- and mu
lti-aspect sentiment classification tasks and demonstrate that the proposed meth
od is able to identify both factual (justifying the ground truth label) and coun
terfactual (countering the ground truth label) rationales consistent with human
rationalization. The code for our method is publicly available.
****
Scalable Bayesian dynamic covariance modeling with variational Wishart and inver
se Wishart processes
Creighton Heaukulani, Mark van der Wilk
We implement gradient-based variational inference routines for Wishart and inver
se Wishart processes, which we apply as Bayesian models for the dynamic, heteros
kedastic covariance matrix of a multivariate time series. The Wishart and invers
e Wishart processes are constructed from i.i.d. Gaussian processes, existing var
iational inference algorithms for which form the basis of our approach. These me
thods are easy to implement as a black-box and scale favorably with the length o
f the time series, however, they fail in the case of the Wishart process, an iss
ue we resolve with a simple modification into an additive white noise parameteri
zation of the model. This modification is also key to implementing a factored va
riant of the construction, allowing inference to additionally scale to high-dime
nsional covariance matrices. Through experimentation, we demonstrate that some (
but not all) model variants outperform multivariate GARCH when forecasting the c
ovariances of returns on financial instruments.
****
Variational Bayesian Decision-making for Continuous Utilities
Tomasz Ku■mierczyk, Joseph Sakaya, Arto Klami
Bayesian decision theory outlines a rigorous framework for making optimal decisi
ons based on maximizing expected utility over a model posterior. However, practi
tioners often do not have access to the full posterior and resort to approximate
 inference strategies. In such cases, taking the eventual decision-making task i
nto account while performing the inference allows for calibrating the posterior
approximation to maximize the utility. We present an automatic pipeline that co-
opts continuous utilities into variational inference algorithms to account for d
ecision-making. We provide practical strategies for approximating and maximizing
 the gain, and empirically demonstrate consistent improvement when calibrating a
pproximations for specific utilities.
****
Optimal Sparsity-Sensitive Bounds for Distributed Mean Estimation
zengfeng Huang, Ziyue Huang, Yilei WANG, Ke Yi
We consider the problem of estimating the mean of a set of vectors, which are st
ored in a distributed system. This is a fundamental task with applications in di
stributed SGD and many other distributed problems, where communication is a main
 bottleneck for scaling up computations. We propose a new sparsity-aware algorit
hm, which improves previous results both theoretically and empirically. The comm

unication cost of our algorithm is characterized by Hoyer's measure of sparseness. Moreover, we prove that the communication cost of our algorithm is information-theoretic optimal up to a constant factor in all sparseness regime. We have also conducted experimental studies, which demonstrate the advantages of our method and confirm our theoretical findings.
****

Search on the Replay Buffer: Bridging Planning and Reinforcement Learning
Ben Eysenbach, Russ R. Salakhutdinov, Sergey Levine
The history of learning for control has been an exciting back and forth between two broad classes of algorithms: planning and reinforcement learning. Planning algorithms effectively reason over long horizons, but assume access to a local policy and distance metric over collision-free paths. Reinforcement learning excels at learning policies and relative values of states, but fails to plan over long horizons. Despite the successes of each method on various tasks, long horizon, sparse reward tasks with high-dimensional observations remain exceedingly challenging for both planning and reinforcement learning algorithms. Frustratingly, these sorts of tasks are potentially the most useful, as they are simple to design (a human only need to provide an example goal state) and avoid injecting bias through reward shaping. We introduce a general-purpose control algorithm that combines the strengths of planning and reinforcement learning to effectively solve these tasks. Our main idea is to decompose the task of reaching a distant goal state into a sequence of easier tasks, each of which corresponds to reaching a particular subgoal. We use goal-conditioned RL to learn a policy to reach each waypoint and to learn a distance metric for search. Using graph search over our replay buffer, we can automatically generate this sequence of subgoals, even in image-based environments. Our algorithm, search on the replay buffer (SoRB), enables agents to solve sparse reward tasks over hundreds of steps, and generalizes substantially better than standard RL algorithms.
****

Minimal Variance Sampling in Stochastic Gradient Boosting
Bulat Ibragimov, Gleb Gusev
Stochastic Gradient Boosting (SGB) is a widely used approach to regularization of boosting models based on decision trees. It was shown that, in many cases, random sampling at each iteration can lead to better generalization performance of the model and can also decrease the learning time. Different sampling approaches were proposed, where probabilities are not uniform, and it is not currently clear which approach is the most effective. In this paper, we formulate the problem of randomization in SGB in terms of optimization of sampling probabilities to maximize the estimation accuracy of split scoring used to train decision trees.This optimization problem has a closed-form nearly optimal solution, and it leads to a new sampling technique, which we call Minimal Variance Sampling (MVS).The method both decreases the number of examples needed for each iteration of boosting and increases the quality of the model significantly as compared to the state-of-the art sampling methods. The superiority of the algorithm was confirmed by introducing MVS as a new default option for subsampling in CatBoost, a gradient boosting library achieving state-of-the-art quality on various machine learning tasks.
****

Transductive Zero-Shot Learning with Visual Structure Constraint
Ziyu Wan, Dongdong Chen, Yan Li, Xingguang Yan, Junge Zhang, Yizhou Yu, Jing Liao
To recognize objects of the unseen classes, most existing Zero-Shot Learning (ZSL) methods first learn a compatible projection function between the common semantic space and the visual space based on the data of source seen classes, then directly apply it to the target unseen classes. However, in real scenarios, the data distribution between the source and target domain might not match well, thus causing the well-known domain shift problem. Based on the observation that visual features of test instances can be separated into different clusters, we propose a new visual structure constraint on class centers for transductive ZSL, to improve the generality of the projection function (\ie alleviate the above domain

shift problem). Specifically, three different strategies (symmetric Chamfer-distance,Bipartite matching distance, and Wasserstein distance) are adopted to align the projected unseen semantic centers and visual cluster centers of test instances. We also propose a new training strategy to handle the real cases where many unrelated images exist in the test dataset, which is not considered in previous methods. Experiments on many widely used datasets demonstrate that the proposed visual structure constraint can bring substantial performance gain consistently and achieve state-of-the-art results.
****
Large Scale Markov Decision Processes with Changing Rewards
Adrian Rivera Cardoso, He Wang, Huan 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.
****
A Unified Framework for Data Poisoning Attack to Graph-based Semi-supervised Learning
Xuanqing Liu, Si Si, Jerry Zhu, Yang Li, 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.
****
Implicit Regularization for Optimal Sparse Recovery
Tomas Vaskevicius, Varun Kanade, Patrick Rebeschini
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.
****
Residual Flows for Invertible Generative Modeling
Ricky T. Q. Chen, Jens Behrmann, David K. Duvenaud, Joern-Henrik Jacobsen
Flow-based generative models parameterize probability distributions through an invertible transformation and can be trained by maximum likelihood. Invertible residual networks provide a flexible family of transformations where only Lipschitz conditions rather than strict architectural constraints are needed for enforcing invertibility. However, prior work trained invertible residual networks for density estimation by relying on biased log-density estimates whose bias increased with the network's expressiveness. We give a tractable unbiased estimate of the log density, and reduce the memory required during training by a factor of ten. Furthermore, we improve invertible residual blocks by proposing the use of activation functions that avoid gradient saturation and generalizing the Lipschitz condition to induced mixed norms. The resulting approach, called Residual Flows, achieves state-of-the-art performance on density estimation amongst flow-based models, and outperforms networks that use coupling blocks at joint generative and discriminative modeling.
****
Copula Multi-label Learning
Weiwei Liu
A formidable challenge in multi-label learning is to model the interdependencies between labels and features. Unfortunately, the statistical properties of existing multi-label dependency modelings are still not well understood. Copulas are a powerful tool for modeling dependence of multivariate data, and achieve great success in a wide range of applications, such as finance, econometrics and systems neuroscience. This inspires us to develop a novel copula multi-label learning paradigm for modeling label and feature dependencies. The copula based paradigm enables to reveal new statistical insights in multi-label learning. In particular, the paper first leverages the kernel trick to construct continuous distribution in the output space, and then estimates our proposed model semiparametricall

y where the copula is modeled parametrically, while the marginal distributions are modeled nonparametrically. Theoretically, we show that our estimator is an unbiased and consistent estimator and follows asymptotically a normal distribution. Moreover, we bound the mean squared error of estimator. The experimental results from various domains validate the superiority of our proposed approach.
****

Adversarial Training and Robustness for Multiple Perturbations
Florian Tramer, Dan Boneh
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.
****

Certainty Equivalence is Efficient for Linear Quadratic Control
Horia Mania, Stephen Tu, Benjamin Recht
We study the performance of the certainty equivalent controller on Linear Quadratic (LQ) control problems with unknown transition dynamics. We show that for both the fully and partially observed settings, the sub-optimality gap between the cost incurred by playing the certainty equivalent controller on the true system and the cost incurred by using the optimal LQ controller enjoys a fast statistical rate, scaling as the square of the parameter error. To the best of our knowledge, our result is the first sub-optimality guarantee in the partially observed Linear Quadratic Gaussian (LQG) setting. Furthermore, in the fully observed Linear Quadratic Regulator (LQR), our result improves upon recent work by Dean et al., who present an algorithm achieving a sub-optimality gap linear in the parameter error. A key part of our analysis relies on perturbation bounds for discrete Riccati equations. We provide two new perturbation bounds, one that expands on an existing result from Konstantinov, and another based on a new elementary proof strategy.
****

Stein Variational Gradient Descent With Matrix-Valued Kernels
Dilin Wang, Ziyang Tang, Chandrajit Bajaj, Qiang Liu
Stein variational gradient descent (SVGD) is a particle-based inference algorithm that leverages gradient information for efficient approximate inference. In this work, we enhance SVGD by leveraging preconditioning matrices, such as the Hessian and Fisher information matrix, to incorporate geometric information into SVGD updates. We achieve this by presenting a generalization of SVGD that replaces the scalar-valued kernels in vanilla SVGD with more general matrix-valued kernels. This yields a significant extension of SVGD, and more importantly, allows us to flexibly incorporate various preconditioning matricesto accelerate the exploration in the probability landscape. Empirical results show that our method outperforms vanilla SVGD and a variety of baseline approaches over a range of real-world Bayesian inference tasks.
****

Differentially Private Bagging: Improved utility and cheaper privacy than subsample-and-aggregate
James Jordon, Jinsung Yoon, Mihaela van der Schaar
Differential Privacy is a popular and well-studied notion of privacy. In the era ofbig data that we are in, privacy concerns are becoming ever more prevalent and thusdifferential privacy is being turned to as one such solution. A popular method forensuring differential privacy of a classifier is known as subsample-and-aggregate,in which the dataset is divided into distinct chunks and a model is learned on eachchunk, after which it is aggregated. This approach allows for easy analysis of themodel on the data and thus differential privacy can be easily applied. In this paper,we extend this approach by dividing the data several times (rather than just once)and learning models on each chunk within each division. The first benefit of thisapproach is the natural improvement of utility by aggregating models trained ona more diverse range of subsets of the data (as demonstrated by the well-knownbagging technique). The second benefit is that, through analysis that we provide inthe paper, we can derive tighter differential privacy gua

rantees when several queriesare made to this mechanism. In order to derive thes
e guarantees, we introducethe upwards and downwards moments accountants and deri
ve bounds for thesemoments accountants in a data-driven fashion. We demonstrate
the improvementsour model makes over standard subsample-and-aggregate in two dat
asets (HeartFailure (private) and UCI Adult (public)).
****
Abstraction based Output Range Analysis for Neural Networks
Pavithra Prabhakar, Zahra Rahimi Afzal
In this paper, we consider the problem of output range analysis for feed-forward
 neural networks. The current approaches reduce the problem to satisfiability an
d optimization solving which are NP-hard problems, and whose computational compl
exity increases with the number of neurons in the network. We present a novel ab
straction technique that constructs a simpler neural network with fewer neurons,
 albeit with interval weights called interval neural network (INN) which over-ap
proximates the output range of the given neural network. We reduce the output ra
nge analysis on the INNs to solving a mixed integer linear programming problem.
Our experimental results highlight the trade-off between the computation time an
d the precision of the computed output range.
****
Paraphrase Generation with Latent Bag of Words
Yao Fu, Yansong Feng, John P. Cunningham
Paraphrase generation is a longstanding important problem in natural language pr
ocessing.
  Recent progress in deep generative models has shown promising results on discr
ete latent variables for text generation.
  Inspired by variational autoencoders with discrete latent structures,
  in this work, we propose a latent bag of words (BOW) model for paraphrase gene
ration.
  We ground the semantics of a discrete latent variable by the target BOW.
  We use this latent variable to build a fully differentiable content planning a
nd surface realization pipeline.
  Specifically, we use source words to predict their neighbors and model the tar
get BOW with a mixture of softmax.
  We use gumbel top-k reparameterization to perform differentiable subset sampli
ng from the predicted BOW distribution.
  We retrieve the sampled word embeddings and use them to augment the decoder an
d guide its generation search space.
  Our latent BOW model not only enhances the decoder, but also exhibits clear in
terpretability.
  We show the model interpretability with regard to (1). unsupervised learning o
f word neighbors (2). the step-by-step generation procedure.
  Extensive experiments demonstrate the model's transparent and effective genera
tion process.
****
Combinatorial Bandits with Relative Feedback
Aadirupa Saha, Aditya Gopalan
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.
****
Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian P
rocesses
Greg Yang
Wide neural networks with random weights and biases are Gaussian processes, as o
bserved by Neal (1995) for shallow networks, and more recently by Lee et al.~(20
18) and Matthews et al.~(2018) for deep fully-connected networks, as well as by
Novak et al.~(2019) and Garriga-Alonso et al.~(2019) for deep convolutional netw
orks.
We show that this Neural Network-Gaussian Process correspondence surprisingly ex

tends to all modern feedforward or recurrent neural networks composed of multila
yer perceptron, RNNs (e.g. LSTMs, GRUs), (nD or graph) convolution, pooling, ski
p connection, attention, batch normalization, and/or layer normalization.
More generally, we introduce a language for expressing neural network computatio
ns, and our result encompasses all such expressible neural networks.
This work serves as a tutorial on the \emph{tensor programs} technique formulate
d in Yang (2019) and elucidates the Gaussian Process results obtained there.
We provide open-source implementations of the Gaussian Process kernels of simple
 RNN, GRU, transformer, and batchnorm+ReLU network at github.com/thegregyang/GP
4A.
Please see our arxiv version for the complete and up-to-date version of this pap
er.
****
An Accelerated Decentralized Stochastic Proximal Algorithm for Finite Sums
Hadrien Hendrikx, Francis Bach, Laurent Massoulié
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.
****
Sample Efficient Active Learning of Causal Trees
Kristjan Greenewald, Dmitriy Katz, Karthikeyan Shanmugam, Sara Magliacane, Murat
 Kocaoglu, Enric Boix Adsera, Guy 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Data Cleansing for Models Trained with SGD
Satoshi Hara, Atsushi Nitanda, Takanori Maehara
Data cleansing is a typical approach used to improve the accuracy of machine lea
rning models, which, however, requires extensive domain knowledge to identify th
e influential instances that affect the models. In this paper, we propose an alg
orithm that can identify influential instances without using any domain knowledg
e. The proposed algorithm automatically cleans the data, which does not require
any of the users' knowledge. Hence, even non-experts can improve the models. The
 existing methods require the loss function to be convex and an optimal model to
 be obtained, which is not always the case in modern machine learning. To overco
me these limitations, we propose a novel approach specifically designed for the
models trained with stochastic gradient descent (SGD). The proposed method infer
s the influential instances by retracing the steps of the SGD while incorporatin
g intermediate models computed in each step. Through experiments, we demonstrate
 that the proposed method can accurately infer the influential instances. Moreov
er, we used MNIST and CIFAR10 to show that the models can be effectively improve
d by removing the influential instances suggested by the proposed method.
****
Universality and individuality in neural dynamics across large populations of re
current networks
Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, David Sussill
o
Many recent studies have employed task-based modeling with recurrent neural netw
orks (RNNs) to infer the computational function of different brain regions. Thes
e models are often assessed by quantitatively comparing the low-dimensional neur
al dynamics of the model and the brain, for example using canonical correlation
analysis (CCA). However, the nature of the detailed neurobiological inferences o
ne can draw from such efforts remains elusive. For example, to what extent does
training neural networks to solve simple tasks, prevalent in neuroscientific stu
dies, uniquely determine the low-dimensional dynamics independent of neural arch
itectures? Or alternatively, are the learned dynamics highly sensitive to diffe
rent neural architectures? Knowing the answer to these questions has strong imp

lications on whether and how to use task-based RNN modeling to understand brain dynamics. To address these foundational questions, we study populations of thous ands of networks of commonly used RNN architectures trained to solve neuroscient ifically motivated tasks and characterize their low-dimensional dynamics via CCA and nonlinear dynamical systems analysis. We find the geometry of the dynamics can be highly sensitive to different network architectures, and further find str iking dissociations between geometric similarity as measured by CCA and network function, yielding a cautionary tale. Moreover, we find that while the geometry of neural dynamics can vary greatly across architectures, the underlying computa tional scaffold: the topological structure of fixed points, transitions between them, limit cycles, and linearized dynamics, often appears {\it universal} acros s all architectures. Overall, this analysis of universality and individuality a cross large populations of RNNs provides a much needed foundation for interpreti ng quantitative measures of dynamical similarity between RNN and brain dynamics.
****
Generating Diverse High-Fidelity Images with VQ-VAE-2
Ali Razavi, Aaron van den Oord, Oriol Vinyals
We explore the use of Vector Quantized Variational AutoEncoder (VQ-VAE) models f or large scale image generation.
To this end, we scale and enhance the autoregressive priors used in VQ-VAE to ge nerate synthetic samples of much higher coherence and fidelity than possible bef ore.
We use simple feed-forward encoder and decoder networks, making our model an att ractive candidate for applications where the encoding and/or decoding speed is c ritical. Additionally, VQ-VAE requires sampling an autoregressive model only in the compressed latent space, which is an order of magnitude faster than samplin g in the pixel space, especially for large images.
We demonstrate that a multi-scale hierarchical organization of VQ-VAE, augmente d with powerful priors over the latent codes, is able to generate samples with q uality that rivals that of state of the art Generative Adversarial Networks on m ultifaceted datasets such as ImageNet, while not suffering from GAN's known shor tcomings such as mode collapse and lack of diversity.
****
When to Trust Your Model: Model-Based Policy Optimization
Michael Janner, Justin Fu, Marvin Zhang, Sergey Levine
Designing effective model-based reinforcement learning algorithms is difficult b ecause the ease of data generation must be weighed against the bias of model-gen erated data. In this paper, we study the role of model usage in policy optimizat ion both theoretically and empirically. We first formulate and analyze a model-b ased reinforcement learning algorithm with a guarantee of monotonic improvement at each step. In practice, this analysis is overly pessimistic and suggests that real off-policy data is always preferable to model-generated on-policy data, bu t we show that an empirical estimate of model generalization can be incorporated into such analysis to justify model usage. Motivated by this analysis, we then demonstrate that a simple procedure of using short model-generated rollouts bran ched from real data has the benefits of more complicated model-based algorithms without the usual pitfalls. In particular, this approach surpasses the sample ef ficiency of prior model-based methods, matches the asymptotic performance of the best model-free algorithms, and scales to horizons that cause other model-based methods to fail entirely.
****
On Making Stochastic Classifiers Deterministic
Andrew Cotter, Maya Gupta, Harikrishna Narasimhan
Stochastic classifiers arise in a number of machine learning problems, and have become especially prominent of late, as they often result from constrained optim ization problems, e.g. for fairness, churn, or custom losses. Despite their util ity, the inherent randomness of stochastic classifiers may cause them to be prob lematic to use in practice for a variety of practical reasons. In this paper, we attempt to answer the theoretical question of how well a stochastic classifier can be approximated by a deterministic one, and compare several different approa

ches, proving lower and upper bounds. We also experimentally investigate the pros and cons of these methods, not only in regard to how successfully each deterministic classifier approximates the original stochastic classifier, but also in terms of how well each addresses the other issues that can make stochastic classifiers undesirable.
****

Blind Super-Resolution Kernel Estimation using an Internal-GAN

Sefi Bell-Kligler, Assaf Shocher, Michal Irani

Super resolution (SR) methods typically assume that the low-resolution (LR) image was downscaled from the unknown high-resolution (HR) image by a fixed `ideal' downscaling kernel (e.g. Bicubic downscaling). However, this is rarely the case in real LR images, in contrast to synthetically generated SR datasets. When the assumed downscaling kernel deviates from the true one, the performance of SR methods significantly deteriorates. This gave rise to Blind-SR - namely, SR when the downscaling kernel (SR-kernel'') is unknown. It was further shown that the true SR-kernel is the one that maximizes the recurrence of patches across scales of the LR image. In this paper we show how this powerful cross-scale recurrence property can be realized using Deep Internal Learning. We introduceKernelGAN'', an image-specific Internal-GAN, which trains solely on the LR test image at test time, and learns its internal distribution of patches. Its Generator is trained to produce a downscaled version of the LR test image, such that its Discriminator cannot distinguish between the patch distribution of the downscaled image, and the patch distribution of the original LR image. The Generator, once trained, constitutes the downscaling operation with the correct image-specific SR-kernel. KernelGAN is fully unsupervised, requires no training data other than the input image itself, and leads to state-of-the-art results in Blind-SR when plugged into existing SR algorithms.
****

Learning to Learn By Self-Critique

Antreas Antoniou, Amos J. Storkey

In few-shot learning, a machine learning system is required to learn from a small set of labelled examples of a specific task, such that it can achieve strong generalization on new unlabelled examples of the same task. Given the limited availability of labelled examples in such tasks, we need to make use of all the information we can. For this reason we propose the use of transductive meta-learning for few shot settings to obtain state-of-the-art few-shot learning.
****

Learning New Tricks From Old Dogs: Multi-Source Transfer Learning From Pre-Trained Networks

Joshua Lee, Prasanna Sattigeri, Gregory Wornell

The advent of deep learning algorithms for mobile devices and sensors has led to a dramatic expansion in the availability and number of systems trained on a wide range of machine learning tasks, creating a host of opportunities and challenges in the realm of transfer learning. Currently, most transfer learning methods require some kind of control over the systems learned, either by enforcing constraints during the source training, or through the use of a joint optimization objective between tasks that requires all data be co-located for training. However, for practical, privacy, or other reasons, in a variety of applications we may have no control over the individual source task training, nor access to source training samples. Instead we only have access to features pre-trained on such data as the output of "black-boxes.'' For such scenarios, we consider the multi-source learning problem of training a classifier using an ensemble of pre-trained neural networks for a set of classes that have not been observed by any of the source networks, and for which we have very few training samples. We show that by using these distributed networks as feature extractors, we can train an effective classifier in a computationally-efficient manner using tools from (nonlinear) maximal correlation analysis. In particular, we develop a method we refer to as maximal correlation weighting (MCW) to build the required target classifier from an appropriate weighting of the feature functions from the source networks. We illustrate the effectiveness of the resulting classifier on datasets deri

ved from the CIFAR-100, Stanford Dogs, and Tiny ImageNet datasets, and, in addition, use the methodology to characterize the relative value of different source tasks in learning a target task.
****

Globally Convergent Newton Methods for Ill-conditioned Generalized Self-concordant Losses
Ulysse Marteau-Ferey, Francis Bach, Alessandro Rudi
In this paper, we study large-scale convex optimization algorithms based on the Newton method applied to regularized generalized self-concordant losses, which include logistic regression and softmax regression. We first prove that our new simple scheme based on a sequence of problems with decreasing regularization parameters is provably globally convergent, that this convergence is linear with a constant factor which scales only logarithmically with the condition number. In the parametric setting, we obtain an algorithm with the same scaling than regular first-order methods but with an improved behavior, in particular in ill-conditioned problems. Second, in the non parametric machine learning setting, we provide an explicit algorithm combining the previous scheme with Nystr\"om projections techniques, and prove that it achieves optimal generalization bounds with a time complexity of order O(n df), a memory complexity of order O(df^2) and no dependence on the condition number, generalizing the results known for least squares regression. Here n is the number of observations and df is the associated degrees of freedom. In particular, this is the first large-scale algorithm to solve logistic and softmax regressions in the non-parametric setting with large condition numbers and theoretical guarantees.
****

Is Deeper Better only when Shallow is Good?
Eran Malach, Shai Shalev-Shwartz
Understanding the power of depth in feed-forward neural networks is an ongoing challenge in the field of deep learning theory. While current works account for the importance of depth for the expressive power of neural-networks, it remains an open question whether these benefits are exploited during a gradient-based optimization process.
In this work we explore the relation between expressivity properties of deep networks and the ability to train them efficiently using gradient-based algorithms. We give a depth separation argument for distributions with fractal structure, showing that they can be expressed efficiently by deep networks, but not with shallow ones.
These distributions have a natural coarse-to-fine structure, and we show that the balance between the coarse and fine details has a crucial effect on whether the optimization process is likely to succeed. We prove that when the distribution is concentrated on the fine details, gradient-based algorithms are likely to fail.
Using this result we prove that, at least in some distributions, the success of learning deep networks depends on whether the distribution can be approximated by shallower networks, and we conjecture that this property holds in general.
****

Variance Reduced Policy Evaluation with Smooth Function Approximation
Hoi-To Wai, Mingyi Hong, Zhuoran Yang, Zhaoran Wang, Kexin 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.
****

k-Means Clustering of Lines for Big Data
Yair Marom, Dan Feldman
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 Leakage from Gradients
Ligeng Zhu, Zhijian Liu, Song Han
Passing gradient is a widely used scheme in modern multi-node learning system (e.g, distributed training, collaborative learning). In a long time, people used to believe that gradients are safe to share: i.e, the training set will not be leaked by gradient sharing. However, in this paper, we show that we can obtain the private training set from the publicly shared gradients. The leaking only takes few gradient steps to process and can obtain the original training set instead of look-alike alternatives. We name this leakage as \textit{deep leakage from gradient} and practically validate the effectiveness of our algorithm on both computer vision and natural language processing tasks. We empirically show that our attack is much stronger than previous approaches and thereby and raise people's awareness to rethink the gradients' safety. We also discuss some possible strategies to defend this deep leakage.
****
Robustness to Adversarial Perturbations in Learning from Incomplete Data
Amir Najafi, Shin-ichi Maeda, Masanori Koyama, Takeru Miyato
What is the role of unlabeled data in an inference problem, when the presumed underlying distribution is adversarially perturbed? To provide a concrete answer to this question, this paper unifies two major learning frameworks: Semi-Supervised Learning (SSL) and Distributionally Robust Learning (DRL). We develop a generalization theory for our framework based on a number of novel complexity measures, such as an adversarial extension of Rademacher complexity and its semi-supervised analogue. Moreover, our analysis is able to quantify the role of unlabeled data in the generalization under a more general condition compared to the existing theoretical works in SSL. Based on our framework, we also present a hybrid of DRL and EM algorithms that has a guaranteed convergence rate. When implemented with deep neural networks, our method shows a comparable performance to those of the state-of-the-art on a number of real-world benchmark datasets.
****
Pure Exploration with Multiple Correct Answers
Rémy Degenne, Wouter M. Koolen
We determine the sample complexity of pure exploration bandit problems with multiple good answers. We derive a lower bound using a new game equilibrium argument. We show how continuity and convexity properties of single-answer problems ensure that the existing Track-and-Stop algorithm has asymptotically optimal sample complexity. However, that convexity is lost when going to the multiple-answer setting. We present a new algorithm which extends Track-and-Stop to the multiple-answer case and has asymptotic sample complexity matching the lower bound.
****
Correlation in Extensive-Form Games: Saddle-Point Formulation and Benchmarks
Gabriele Farina, Chun Kai Ling, Fei Fang, Tuomas Sandholm
While Nash equilibrium in extensive-form games is well understood, very little is known about the properties of extensive-form correlated equilibrium (EFCE), both from a behavioral and from a computational point of view. In this setting, the strategic behavior of players is complemented by an external device that privately recommends moves to agents as the game progresses; players are free to deviate at any time, but will then not receive future recommendations. Our contributions are threefold. First, we show that an EFCE
can be formulated as the solution to a bilinear saddle-point problem. To showcase how this novel formulation can inspire new algorithms to compute EFCEs, we propose a simple subgradient descent method which exploits this formulation and structural properties of EFCEs. Our method has better scalability than the prior approach based on linear programming. Second, we propose two benchmark games, which we hope will serve as the basis for future evaluation of EFCE solvers. These games were chosen so as to cover two natural application domains for EFCE: conflict resolution via a mediator, and bargaining and negotiation. Third, we document the qualitative behavior of EFCE in our proposed games. We show that the social-welfare-maximizing equilibria in these games are highly nontrivial and exhibit surprisingly subtle sequential behavior that so far has not received attention i

n the literature.
****
The Thermodynamic Variational Objective
Vaden Masrani, Tuan Anh Le, Frank Wood
We introduce the thermodynamic variational objective (TVO) for learning in both continuous and discrete deep generative models. The TVO arises from a key connection between variational inference and thermodynamic integration that results in a tighter lower bound to the log marginal likelihood than the standard variational evidence lower bound (ELBO) while remaining as broadly applicable. We provide a computationally efficient gradient estimator for the TVO that applies to continuous, discrete, and non-reparameterizable distributions and show that the objective functions used in variational inference, variational autoencoders, wake sleep, and inference compilation are all special cases of the TVO. We use the TVO to learn both discrete and continuous deep generative models and empirically demonstrate state of the art model and inference network learning.
****
Sampling Sketches for Concave Sublinear Functions of Frequencies
Edith Cohen, Ofir Geri
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 Interpretable Kernel Dimensionality Reduction
Chieh Wu, Jared Miller, Yale Chang, Mario Sznaier, Jennifer Dy
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 Imbalanced Datasets with Label-Distribution-Aware Margin Loss
Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, Tengyu Ma
Deep learning algorithms can fare poorly when the training dataset suffers from heavy class-imbalance but the testing criterion requires good generalization on less frequent classes. We design two novel methods to improve performance in such scenarios. First, we propose a theoretically-principled label-distribution-aware margin (LDAM) loss motivated by minimizing a margin-based generalization bound. This loss replaces the standard cross-entropy objective during training and can be applied with prior strategies for training with class-imbalance such as re-weighting or re-sampling. Second, we propose a simple, yet effective, training schedule that defers re-weighting until after the initial stage, allowing the model to learn an initial representation while avoiding some of the complications associated with re-weighting or re-sampling. We test our methods on several benchmark vision tasks including the real-world imbalanced dataset iNaturalist 2018. Our experiments show that either of these methods alone can already improve over existing techniques and their combination achieves even better performance gains.
****
Multivariate Triangular Quantile Maps for Novelty Detection
Jingjing Wang, Sun Sun, Yaoliang Yu
Novelty detection, a fundamental task in machine learning, has drawn a lot of recent attention due to its wide-ranging applications and the rise of neural approaches. In this work, we present a general framework for neural novelty detection that centers around a multivariate extension of the univariate quantile function. Our framework unifies and extends many classical and recent novelty detection algorithms, and opens the way to exploit recent advances in flow-based neural density estimation. We adapt the multiple gradient descent algorithm to obtain the first efficient end-to-end implementation of our framework that is free of tuning hyperparameters. Extensive experiments over a number of real datasets confirm the efficacy of our proposed method against state-of-the-art alternatives.

****
Gradient-based Adaptive Markov Chain Monte Carlo
Michalis Titsias, Petros Dellaportas
We introduce a gradient-based learning method to automatically adapt Markov chain Monte Carlo (MCMC) proposal distributions to intractable targets. We define a maximum entropy regularised objective function, referred to as generalised speed measure, which can be robustly optimised over the parameters of the proposal distribution by applying stochastic gradient optimisation. An advantage of our method compared to traditional adaptive MCMC methods is that the adaptation occurs even when candidate state values are rejected. This is a highly desirable property of any adaptation strategy because the adaptation starts in early iterations even if the initial proposal distribution is far from optimum. We apply the framework for learning multivariate random walk Metropolis and Metropolis-adjusted Langevin proposals with full covariance matrices, and provide empirical evidence that our method can outperform other MCMC algorithms, including Hamiltonian Monte Carlo schemes.
****
Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers
Zeyuan Allen-Zhu, Yuanzhi Li, Yingyu Liang
The fundamental learning theory behind neural networks remains largely open. What classes of functions can neural networks actually learn? Why doesn't the trained network overfit when it is overparameterized?
****
Online Forecasting of Total-Variation-bounded Sequences
Dheeraj Baby, Yu-Xiang 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.
****
Approximation Ratios of Graph Neural Networks for Combinatorial Problems
Ryoma Sato, Makoto Yamada, Hisashi Kashima
In this paper, from a theoretical perspective, we study how powerful graph neural networks (GNNs) can be for learning approximation algorithms for combinatorial problems.
To this end, we first establish a new class of GNNs that can solve a strictly wider variety of problems than existing GNNs. Then, we bridge the gap between GNN theory and the theory of distributed local algorithms. We theoretically demonstrate that the most powerful GNN can learn approximation algorithms for the minimum dominating set problem and the minimum vertex cover problem with some approximation ratios with the aid of the theory of distributed local algorithms. We also show that most of the existing GNNs such as GIN, GAT, GCN, and GraphSAGE cannot perform better than with these ratios. This paper is the first to elucidate approximation ratios of GNNs for combinatorial problems. Furthermore, we prove that adding coloring or weak-coloring to each node feature improves these approximation ratios. This indicates that preprocessing and feature engineering theoretically strengthen model capabilities.
****
Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video
Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, Ian Reid
Recent work has shown that CNN-based depth and ego-motion estimators can be learned using unlabelled monocular videos. However, the performance is limited by unidentified moving objects that violate the underlying static scene assumption in geometric image reconstruction. More significantly, due to lack of proper constraints, networks output scale-inconsistent results over different samples, i.e., the ego-motion network cannot provide full camera trajectories over a long video sequence because of the per-frame scale ambiguity. This paper tackles these challenges by proposing a geometry consistency loss for scale-consistent predictio

ns and an induced self-discovered mask for handling moving objects and occlusions. Since we do not leverage multi-task learning like recent works, our framework is much simpler and more efficient. Comprehensive evaluation results demonstrate that our depth estimator achieves the state-of-the-art performance on the KITTI dataset. Moreover, we show that our ego-motion network is able to predict a globally scale-consistent camera trajectory for long video sequences, and the resulting visual odometry accuracy is competitive with the recent model that is trained using stereo videos. To the best of our knowledge, this is the first work to show that deep networks trained using unlabelled monocular videos can predict globally scale-consistent camera trajectories over a long video sequence.
****
Variational Denoising Network: Toward Blind Noise Modeling and Removal
Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, Lei Zhang
Blind image denoising is an important yet very challenging problem in computer vision due to the complicated acquisition process of real images. In this work we
propose a new variational inference method, which integrates both noise estimation and image denoising into a unique Bayesian framework, for blind image denoising. Specifically, an approximate posterior, parameterized by deep neural networks, is presented by taking the intrinsic clean image and noise variances as latent variables conditioned on the input noisy image. This posterior provides explicit parametric forms for all its involved hyper-parameters, and thus can be easily implemented for blind image denoising with automatic noise estimation for the test noisy image. On one hand, as other data-driven deep learning methods, our method, namely variational denoising network (VDN), can perform denoising efficiently due to its explicit form of posterior expression. On the other hand, VDN inherits the advantages of traditional model-driven approaches, especially the good generalization capability of generative models. VDN has good interpretability and can be flexibly utilized to estimate and remove complicated non-i.i.d. noise collected in real scenarios. Comprehensive experiments are performed to substantiate the superiority of our method in blind image denoising.
****
Multi-task Learning for Aggregated Data using Gaussian Processes
Fariba Yousefi, Michael T. Smith, Mauricio Álvarez
Aggregated data is commonplace in areas such as epidemiology and demography. For example, census data for a population is usually given as averages defined over time periods or spatial resolutions (cities, regions or countries). In this paper, we present a novel multi-task learning model based on Gaussian processes for joint learning of variables that have been aggregated at different input scales. Our model represents each task as the linear combination of the realizations of latent processes that are integrated at a different scale per task. We are then able to compute the cross-covariance between the different tasks either analytically or numerically. We also allow each task to have a potentially different likelihood model and provide a variational lower bound that can be optimised in a stochastic fashion making our model suitable for larger datasets. We show examples of the model in a synthetic example, a fertility dataset and an air pollution prediction application.
****
Keeping Your Distance: Solving Sparse Reward Tasks Using Self-Balancing Shaped Rewards
Alexander Trott, Stephan Zheng, Caiming Xiong, Richard Socher
While using shaped rewards can be beneficial when solving sparse reward tasks, their successful application often requires careful engineering and is problem specific. For instance, in tasks where the agent must achieve some goal state, simple distance-to-goal reward shaping often fails, as it renders learning vulnerable to local optima. We introduce a simple and effective model-free method to learn from shaped distance-to-goal rewards on tasks where success depends on reaching a goal state. Our method introduces an auxiliary distance-based reward based on pairs of rollouts to encourage diverse exploration. This approach effectively prevents learning dynamics from stabilizing around local optima induced by t

he naive distance-to-goal reward shaping and enables policies to efficiently solve sparse reward tasks. Our augmented objective does not require any additional reward engineering or domain expertise to implement and converges to the original sparse objective as the agent learns to solve the task. We demonstrate that our method successfully solves a variety of hard-exploration tasks (including maze navigation and 3D construction in a Minecraft environment), where naive distance-based reward shaping otherwise fails, and intrinsic curiosity and reward relabeling strategies exhibit poor performance.
****

Efficient characterization of electrically evoked responses for neural interfaces
Nishal Shah, Sasidhar Madugula, Pawel Hottowy, Alexander Sher, Alan Litke, Liam Paninski, E.J. Chichilnisky
Future neural interfaces will read and write population neural activity with high spatial and temporal resolution, for diverse applications. For example, an artificial retina may restore vision to the blind by electrically stimulating retinal ganglion cells. Such devices must tune their function, based on stimulating and recording, to match the function of the circuit. However, existing methods for characterizing the neural interface scale poorly with the number of electrodes, limiting their practical applicability. This work tests the idea that using prior information from previous experiments and closed-loop measurements may greatly increase the efficiency of the neural interface. Large-scale, high-density electrical recording and stimulation in primate retina were used as a lab prototype for an artificial retina. Three key calibration steps were optimized: spike sorting in the presence of stimulation artifacts, response modeling, and adaptive stimulation. For spike sorting, exploiting the similarity of electrical artifact across electrodes and experiments substantially reduced the number of required measurements. For response modeling, a joint model that captures the inverse relationship between recorded spike amplitude and electrical stimulation threshold from previously recorded retinas resulted in greater consistency and efficiency. For adaptive stimulation, choosing which electrodes to stimulate based on probability estimates from previous measurements improved efficiency. Similar improvements resulted from using either non-adaptive stimulation with a joint model across cells, or adaptive stimulation with an independent model for each cell. Finally, image reconstruction revealed that these improvements may translate to improved performance of an artificial retina.
****

The Synthesis of XNOR Recurrent Neural Networks with Stochastic Logic
Arash Ardakani, Zhengyun Ji, Amir Ardakani, Warren Gross
The emergence of XNOR networks seek to reduce the model size and computational cost of neural networks for their deployment on specialized hardware requiring real-time processes with limited hardware resources. In XNOR networks, both weights and activations are binary, bringing great benefits to specialized hardware by replacing expensive multiplications with simple XNOR operations. Although XNOR convolutional and fully-connected neural networks have been successfully developed during the past few years, there is no XNOR network implementing commonly-used variants of recurrent neural networks such as long short-term memories (LSTMs). The main computational core of LSTMs involves vector-matrix multiplications followed by a set of non-linear functions and element-wise multiplications to obtain the gate activations and state vectors, respectively. Several previous attempts on quantization of LSTMs only focused on quantization of the vector-matrix multiplications in LSTMs while retaining the element-wise multiplications in full precision. In this paper, we propose a method that converts all the multiplications in LSTMs to XNOR operations using stochastic computing. To this end, we introduce a weighted finite-state machine and its synthesis method to approximate the non-linear functions used in LSTMs on stochastic bit streams. Experimental results show that the proposed XNOR LSTMs reduce the computational complexity of their quantized counterparts by a factor of 86x without any sacrifice on latency while achieving a better accuracy across various temporal tasks.
****

HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models
Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F. Fei-Fei, Michael Bernstein
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.
****
McDiarmid-Type Inequalities for Graph-Dependent Variables and Stability Bounds
Rui (Ray) Zhang, Xingwu Liu, Yuyi Wang, Liwei Wang
A crucial assumption in most statistical learning theory is that samples are independently and identically distributed (i.i.d.). However, for many real applications, the i.i.d. assumption does not hold. We consider learning problems in which examples are dependent and their dependency relation is characterized by a graph. To establish algorithm-dependent generalization theory for learning with non-i.i.d. data, we first prove novel McDiarmid-type concentration inequalities for Lipschitz functions of graph-dependent random variables. We show that concentration relies on the forest complexity of the graph, which characterizes the strength of the dependency. We demonstrate that for many types of dependent data, the forest complexity is small and thus implies good concentration. Based on our new inequalities we are able to build stability bounds for learning from graph-dependent data.
****
Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices
Santosh Vempala, Andre Wibisono
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.
****
Are sample means in multi-armed bandits positively or negatively biased?
Jaehyeok Shin, Aaditya Ramdas, Alessandro Rinaldo
It is well known that in stochastic multi-armed bandits (MAB), the sample mean of an arm is typically not an unbiased estimator of its true mean. In this paper, we decouple three different sources of this selection bias: adaptive \emph{sampling} of arms, adaptive \emph{stopping} of the experiment, and adaptively \emph{choosing} which arm to study. Through a new notion called ``optimism'' that captures certain natural monotonic behaviors of algorithms, we provide a clean and unified analysis of how optimistic rules affect the sign of the bias. The main takeaway message is that optimistic sampling induces a negative bias, but optimistic stopping and optimistic choosing both induce a positive bias. These results are derived in a general stochastic MAB setup that is entirely agnostic to the final aim of the experiment (regret minimization or best-arm identification or anything else). We provide examples of optimistic rules of each type, demonstrate that simulations confirm our theoretical predictions, and pose some natural but hard open problems.
****
The Landscape of Non-convex Empirical Risk with Degenerate Population Risk
Shuang Li, Gongguo Tang, Michael B. Wakin
The landscape of empirical risk has been widely studied in a series of machine learning problems, including low-rank matrix factorization, matrix sensing, matrix completion, and phase retrieval. In this work, we focus on the situation where the corresponding population risk is a degenerate non-convex loss function, namely, the Hessian of the population risk can have zero eigenvalues. Instead of analyzing the non-convex empirical risk directly, we first study the landscape of the corresponding population risk, which is usually easier to characterize, and then build a connection between the landscape of the empirical risk and its population risk. In particular, we establish a correspondence between the critical points of the empirical risk and its population risk without the strongly Morse assumption, which is required in existing literature but not satisfied in degener

ate scenarios. We also apply the theory to matrix sensing and phase retrieval to demonstrate how to infer the landscape of empirical risk from that of the corresponding population risk.
****
Hybrid 8-bit Floating Point (HFP8) Training and Inference for Deep Neural Networks
Xiao Sun, Jungwook Choi, Chia-Yu Chen, Naigang Wang, Swagath Venkataramani, Vijayalakshmi (Viji) Srinivasan, Xiaodong Cui, Wei Zhang, Kailash Gopalakrishnan
Reducing the numerical precision of data and computation is extremely effective in accelerating deep learning training workloads. Towards this end, 8-bit floating point representations (FP8) were recently proposed for DNN training. However, its applicability was demonstrated on a few selected models only and significant degradation is observed when popular networks such as MobileNet and Transformer are trained using FP8. This degradation is due to the inherent precision requirement difference in the forward and backward passes of DNN training. Using theoretical insights, we propose a hybrid FP8 (HFP8) format and DNN end-to-end distributed training procedure. We demonstrate, using HFP8, the successful training of deep learning models across a whole spectrum of applications including Image Classification, Object Detection, Language and Speech without accuracy degradation. Finally, we demonstrate that, by using the new 8 bit format, we can directly quantize a pre-trained model down to 8-bits without losing accuracy by simply fine-tuning batch normalization statistics. These novel techniques enable a new generations of 8-bit hardware that are robust for building and deploying neural network models.
****
Are deep ResNets provably better than linear predictors?
Chulhee Yun, Suvrit Sra, Ali Jadbabaie
Recent results in the literature indicate that a residual network (ResNet) composed of a single residual block outperforms linear predictors, in the sense that all local minima in its optimization landscape are at least as good as the best linear predictor. However, these results are limited to a single residual block (i.e., shallow ResNets), instead of the deep ResNets composed of multiple residual blocks. We take a step towards extending this result to deep ResNets. We start by two motivating examples. First, we show that there exist datasets for which all local minima of a fully-connected ReLU network are no better than the best linear predictor, whereas a ResNet has strictly better local minima. Second, we show that even at the global minimum, the representation obtained from the residual block outputs of a 2-block ResNet do not necessarily improve monotonically over subsequent blocks, which highlights a fundamental difficulty in analyzing deep ResNets. Our main theorem on deep ResNets shows under simple geometric conditions that, any critical point in the optimization landscape is either (i) at least as good as the best linear predictor; or (ii) the Hessian at this critical point has a strictly negative eigenvalue. Notably, our theorem shows that a chain of multiple skip-connections can improve the optimization landscape, whereas existing results study direct skip-connections to the last hidden layer or output layer. Finally, we complement our results by showing benign properties of the "near-identity regions" of deep ResNets, showing depth-independent upper bounds for the risk attained at critical points as well as the Rademacher complexity.
****
E2-Train: Training State-of-the-art CNNs with Over 80% Energy Savings
Yue Wang, Ziyu Jiang, Xiaohan Chen, Pengfei Xu, Yang Zhao, Yingyan Lin, Zhangyang Wang
Convolutional neural networks (CNNs) have been increasingly deployed to edge devices. Hence, many efforts have been made towards efficient CNN inference on resource-constrained platforms. This paper attempts to explore an orthogonal direction: how to conduct more energy-efficient training of CNNs, so as to enable on-device training? We strive to reduce the energy cost during training, by dropping unnecessary computations, from three complementary levels: stochastic mini-batch dropping on the data level; selective layer update on the model level; and sign prediction for low-cost, low-precision back-propagation, on the algorithm level

. Extensive simulations and ablation studies, with real energy measurements from an FPGA board, confirm the superiority of our proposed strategies and demonstrate remarkable energy savings for training. For example, when training ResNet-74 on CIFAR-10, we achieve aggressive energy savings of >90% and >60%, while incurring a top-1 accuracy loss of only about 2% and 1.2%, respectively. When training ResNet-110 on CIFAR-100, an over 84% training energy saving is achieved without degrading inference accuracy.
****
Graph Neural Tangent Kernel: Fusing Graph Neural Networks with Graph Kernels
Simon S. Du, Kangcheng Hou, Russ R. Salakhutdinov, Barnabas Poczos, Ruosong Wang, Keyulu Xu
While graph kernels (GKs) are easy to train and enjoy provable theoretical guarantees, their practical performances are limited by their expressive power, as the kernel function often depends on hand-crafted combinatorial features of graphs. Compared to graph kernels, graph neural networks (GNNs) usually achieve better practical performance, as GNNs use multi-layer architectures and non-linear activation functions to extract high-order information of graphs as features. However, due to the large number of hyper-parameters and the non-convex nature of the training procedure, GNNs are harder to train. Theoretical guarantees of GNNs are also not well-understood. Furthermore, the expressive power of GNNs scales with the number of parameters, and thus it is hard to exploit the full power of GNNs when computing resources are limited. The current paper presents a new class of graph kernels, Graph Neural Tangent Kernels (GNTKs), which correspond to \emph{infinitely wide} multi-layer GNNs trained by gradient descent. GNTKs enjoy the full expressive power of GNNs and inherit advantages of GKs. Theoretically, we show GNTKs provably learn a class of smooth functions on graphs. Empirically, we test GNTKs on graph classification datasets and show they achieve strong performance.
****
Privacy-Preserving Q-Learning with Functional Noise in Continuous Spaces
Baoxiang Wang, Nidhi Hegde
We consider differentially private algorithms for reinforcement learning in continuous spaces, such that neighboring reward functions are indistinguishable. This protects the reward information from being exploited by methods such as inverse reinforcement learning. Existing studies that guarantee differential privacy are not extendable to infinite state spaces, as the noise level to ensure privacy will scale accordingly to infinity. Our aim is to protect the value function approximator, without regard to the number of states queried to the function. It is achieved by adding functional noise to the value function iteratively in the training. We show rigorous privacy guarantees by a series of analyses on the kernel of the noise space, the probabilistic bound of such noise samples, and the composition over the iterations. We gain insight into the utility analysis by proving the algorithm's approximate optimality when the state space is discrete. Experiments corroborate our theoretical findings and show improvement over existing approaches.
****
Learning Data Manipulation for Augmentation and Weighting
Zhiting Hu, Bowen Tan, Russ R. Salakhutdinov, Tom M. Mitchell, Eric P. Xing
Manipulating data, such as weighting data examples or augmenting with new instances, has been increasingly used to improve model training. Previous work has studied various rule- or learning-based approaches designed for specific types of data manipulation. In this work, we propose a new method that supports learning different manipulation schemes with the same gradient-based algorithm. Our approach builds upon a recent connection of supervised learning and reinforcement learning (RL), and adapts an off-the-shelf reward learning algorithm from RL for joint data manipulation learning and model training. Different parameterization of the ``data reward'' function instantiates different manipulation schemes. We showcase data augmentation that learns a text transformation network, and data weighting that dynamically adapts the data sample importance. Experiments show the resulting algorithms significantly improve the image and text classification perf

ormance in low data regime and class-imbalance problems.
****

Hyperparameter Learning via Distributional Transfer
Ho Chung Law, Peilin Zhao, Leung Sing Chan, Junzhou Huang, Dino Sejdinovic
Bayesian optimisation is a popular technique for hyperparameter learning but typically requires initial exploration even in cases where similar prior tasks have been solved. We propose to transfer information across tasks using learnt representations of training datasets used in those tasks. This results in a joint Gaussian process model on hyperparameters and data representations. Representations make use of the framework of distribution embeddings into reproducing kernel Hilbert spaces. The developed method has a faster convergence compared to existing baselines, in some cases requiring only a few evaluations of the target objective.
****

Levenshtein Transformer
Jiatao Gu, Changhan Wang, Junbo Zhao
Modern neural sequence generation models are built to either generate tokens step-by-step from scratch or (iteratively) modify a sequence of tokens bounded by a fixed length.   In this work, we develop Levenshtein Transformer, a new partially autoregressive model devised for more flexible and amenable sequence generation. Unlike previous approaches, the basic operations of our model are insertion and deletion. The combination of them facilitates not only generation but also sequence refinement allowing dynamic length changes. We also propose a set of new training techniques dedicated at them, effectively exploiting one as the other's learning signal thanks to their complementary nature. Experiments applying the proposed model achieve comparable or even better performance with much-improved efficiency on both generation (e.g. machine translation, text summarization) and refinement tasks (e.g. automatic post-editing). We further confirm the flexibility of our model by showing a Levenshtein Transformer trained by machine translation can straightforwardly be used for automatic post-editing.
****

Learning Perceptual Inference by Contrasting
Chi Zhang, Baoxiong Jia, Feng Gao, Yixin Zhu, HongJing Lu, Song-Chun Zhu
“Thinking in pictures,” [1] i.e., spatial-temporal reasoning, effortless and instantaneous for humans, is believed to be a significant ability to perform logical induction and a crucial factor in the intellectual history of technology development. Modern Artificial Intelligence (AI), fueled by massive datasets, deeper models, and mighty computation, has come to a stage where (super-)human-level performances are observed in certain specific tasks. However, current AI's ability in “thinking in pictures” is still far lacking behind. In this work, we study how to improve machines' reasoning ability on one challenging task of this kind: Raven's Progressive Matrices (RPM). Specifically, we borrow the very idea of “contrast effects” from the field of psychology, cognition, and education to design and train a permutation-invariant model. Inspired by cognitive studies, we equip our model with a simple inference module that is jointly trained with the perception backbone. Combining all the elements, we propose the Contrastive Perceptual Inference network (CoPINet) and empirically demonstrate that CoPINet sets the new state-of-the-art for permutation-invariant models on two major datasets. We conclude that spatial-temporal reasoning depends on envisaging the possibilities consistent with the relations between objects and can be solved from pixel-level inputs.
****

Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting
Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, Xifeng Yan
Time series forecasting is an important problem across many domains, including predictions of solar plant energy output, electricity consumption, and traffic jam situation. In this paper, we propose to tackle such forecasting problem with Transformer. Although impressed by its performance in our preliminary study, we f

ound its two major weaknesses: (1) locality-agnostics: the point-wise dot- produ
ct self-attention in canonical Transformer architecture is insensitive to local
context, which can make the model prone to anomalies in time series; (2) memory
bottleneck: space complexity of canonical Transformer grows quadratically with s
equence length L, making directly modeling long time series infeasible. In order
 to solve these two issues, we first propose convolutional self-attention by pro
ducing queries and keys with causal convolution so that local context can be bet
ter incorporated into attention mechanism. Then, we propose LogSparse Transforme
r with only O(L(log L)^2) memory cost, improving forecasting accuracy for time s
eries with fine granularity and strong long-term dependencies under constrained
memory budget. Our experiments on both synthetic data and real- world datasets s
how that it compares favorably to the state-of-the-art.
****
Multi-View Reinforcement Learning
Minne Li, Lisheng Wu, Jun WANG, Haitham Bou Ammar
This paper is concerned with multi-view reinforcement learning (MVRL), which all
ows for decision making when agents share common dynamics but adhere to differen
t observation models. We define the MVRL framework by extending partially observ
able Markov decision processes (POMDPs) to support more than one observation mod
el and propose two solution methods through observation augmentation and cross-v
iew policy transfer. We empirically evaluate our method and demonstrate its effe
ctiveness in a variety of environments. Specifically, we show reductions in samp
le complexities and computational time for acquiring policies that handle multi-
view environments.
****
Updates of Equilibrium Prop Match Gradients of Backprop Through Time in an RNN w
ith Static Input
Maxence Ernoult, Julie Grollier, Damien Querlioz, Yoshua Bengio, Benjamin Scelli
er
Equilibrium Propagation (EP) is a biologically inspired learning algorithm for
convergent recurrent neural networks, i.e. RNNs that are fed by a static input x
 and
settle to a steady state. Training convergent RNNs consists in adjusting the wei
ghts
until the steady state of output neurons coincides with a target y. Convergent R
NNs
can also be trained with the more conventional Backpropagation Through Time
(BPTT) algorithm. In its original formulation EP was described in the case of
real-time neuronal dynamics, which is computationally costly. In this work, we
introduce a discrete-time version of EP with simplified equations and with reduc
ed
simulation time, bringing EP closer to practical machine learning tasks. We firs
t
prove theoretically, as well as numerically that the neural and weight updates o
f EP,
computed by forward-time dynamics, are step-by-step equal to the ones obtained b
y
BPTT, with gradients computed backward in time. The equality is strict when the
transition function of the dynamics derives from a primitive function and the st
eady
state is maintained long enough. We then show for more standard discrete-time
neural network dynamics that the same property is approximately respected and
we subsequently demonstrate training with EP with equivalent performance to
BPTT. In particular, we define the first convolutional architecture trained with
 EP
achieving ~ 1% test error on MNIST, which is the lowest error reported with EP.
These results can guide the development of deep neural networks trained with EP.
****
Toward a Characterization of Loss Functions for Distribution Learning
Nika Haghtalab, Cameron Musco, Bo Waggoner

In this work we study loss functions for learning and evaluating probability dis tributions over large discrete domains. Unlike classification or regression whe re a wide variety of loss functions are used, in the distribution learning and d ensity estimation literature, very few losses outside the dominant \emph{log los s} are applied. We aim to understand this fact, taking an axiomatic approach to the design of loss functions for distributions. We start by proposing a set of d esirable criteria that any good loss function should satisfy. Intuitively, these criteria require that the loss function faithfully evaluates a candidate distri bution, both in expectation and when estimated on a few samples. Interestingly, we observe that \emph{no loss function} possesses all of these criteria. However , one can circumvent this issue by introducing a natural restriction on the set of candidate distributions. Specifically, we require that candidates are \emph{ calibrated} with respect to the target distribution, i.e., they may contain less information than the target but otherwise do not significantly distort the trut h. We show that, after restricting to this set of distributions, the log loss an d a large variety of other losses satisfy the desired criteria. These results pa ve the way for future investigations of distribution learning that look beyond t he log loss, choosing a loss function based on application or domain need.
****
Can SGD Learn Recurrent Neural Networks with Provable Generalization?
Zeyuan Allen-Zhu, Yuanzhi Li
Recurrent Neural Networks (RNNs) are among the most popular models in sequential data analysis. Yet, in the foundational PAC learning language, what concept cla ss can it learn? Moreover, how can the same recurrent unit simultaneously learn functions from different input tokens to different output tokens, without affect ing each other?
Existing generalization bounds for RNN scale exponentially with the input length , significantly limiting their practical implications.
****
Image Captioning: Transforming Objects into Words
Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares
Image captioning models typically follow an encoder-decoder architecture which u ses abstract image feature vectors as input to the encoder.
One of the most successful algorithms uses feature vectors extracted from the re gion proposals obtained from an object detector. In this work we introduce the O bject Relation Transformer, that builds upon this approach by explicitly incorpo rating information about the spatial relationship between input detected objects through geometric attention. Quantitative and qualitative results demonstrate t he importance of such geometric attention for image captioning, leading to impro vements on all common captioning metrics on the MS-COCO dataset. Code is availab le at https://github.com/yahoo/objectrelationtransformer .
****
MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis
Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, Aaron C. Courville
Previous works (Donahue et al., 2018a; Engel et al., 2019a) have found that gene rating coherent raw audio waveforms with GANs is challenging. In this paper, we show that it is possible to train GANs reliably to generate high quality coheren t waveforms by introducing a set of architectural changes and simple training te chniques. Subjective evaluation metric (Mean Opinion Score, or MOS) shows the ef fectiveness of the proposed approach for high quality mel-spectrogram inversion. To establish the generality of the proposed techniques, we show qualitative res ults of our model in speech synthesis, music domain translation and unconditiona l music synthesis. We evaluate the various components of the model through ablat ion studies and suggest a set of guidelines to design general purpose discrimina tors and generators for conditional sequence synthesis tasks. Our model is non-a utoregressive, fully convolutional, with significantly fewer parameters than com peting models and generalizes to unseen speakers for mel-spectrogram inversion. Our pytorch implementation runs at more than 100x faster than realtime on GTX 10 80Ti GPU and more than 2x faster than real-time on CPU, without any hardware spe

cific optimization tricks.
****
Deliberative Explanations: visualizing network insecurities
Pei Wang, Nuno Nvasconcelos
A new approach to explainable AI, denoted {\it deliberative explanations,\/} is proposed. Deliberative explanations are a visualization technique that aims to go beyond the simple visualization of the image regions (or, more generally, input variables) responsible for a network prediction. Instead, they aim to expose the deliberations carried by the network to arrive at that prediction, by uncovering the insecurities of the network about the latter. The explanation consists of a list of insecurities, each composed of 1) an image region (more generally, a set of input variables), and 2) an ambiguity formed by the pair of classes responsible for the network uncertainty about the region. Since insecurity detection requires quantifying the difficulty of network predictions, deliberative explanations combine ideas from the literatures on visual explanations and assessment of classification difficulty. More specifically, the proposed implementation combines attributions with respect to both class predictions and a difficulty score. An evaluation protocol that leverages object recognition (CUB200) and scene classification (ADE20K) datasets that combine part and attribute annotations is also introduced to evaluate the accuracy of deliberative explanations. Finally, an experimental evaluation shows that the most accurate explanations are achieved by combining non self-referential difficulty scores and second-order attributions. The resulting insecurities are shown to correlate with regions of attributes that are shared by different classes. Since these regions are also ambiguous for humans, deliberative explanations are intuitive, suggesting that the deliberative process of modern networks correlates with human reasoning.
****
Uncoupled Regression from Pairwise Comparison Data
Liyuan Xu, Junya Honda, Gang Niu, Masashi Sugiyama
Uncoupled regression is the problem to learn a model from unlabeled data and the set of target values while the correspondence between them is unknown. Such a situation arises in predicting anonymized targets that involve sensitive information, e.g., one's annual income. Since existing methods for uncoupled regression often require strong assumptions on the true target function, and thus, their range of applications is limited, we introduce a novel framework that does not require such assumptions in this paper. Our key idea is to utilize \emph{pairwise comparison data, which consists of pairs of unlabeled data that we know which one has a larger target value. Such pairwise comparison data is easy to collect, as typically discussed in the learning-to-rank scenario, and does not break the anonymity of data. We propose two practical methods for uncoupled regression from pairwise comparison data and show that the learned regression model converges to the optimal model with the optimal parametric convergence rate when the target variable distributes uniformly. Moreover, we empirically show that for linear models the proposed methods are comparable to ordinary supervised regression with labeled data.
****
No-Regret Learning in Unknown Games with Correlated Payoffs
Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, Andreas Krause
We consider the problem of learning to play a repeated multi-agent game with an unknown reward function. Single player online learning algorithms attain strong regret bounds when provided with full information feedback, which unfortunately is unavailable in many real-world scenarios. Bandit feedback alone, i.e., observing outcomes only for the selected action, yields substantially worse performance. In this paper, we consider a natural model where, besides a noisy measuremen

t of the obtained reward, the player can also observe the opponents' actions. Th
is feedback model, together with a regularity assumption on the reward function,
 allows us to exploit the correlations among different game outcomes by means of
 Gaussian processes (GPs). We propose a novel confidence-bound based bandit algo
rithm GP-MW, which utilizes the GP model for the reward function and runs a mult
iplicative weight (MW) method. We obtain novel kernel-dependent regret bounds th
at are comparable to the known bounds in the full information setting, while sub
stantially improving upon the existing bandit results. We experimentally demonst
rate the effectiveness of GP-MW in random matrix games, as well as real-world p
roblems of traffic routing and movie recommendation. In our experiments, GP-MW c
onsistently outperforms several baselines, while its performance is often compar
able to methods that have access to full information feedback.
****
Pareto Multi-Task Learning
Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, Sam Kwong
Multi-task learning is a powerful method for solving multiple correlated tasks s
imultaneously. However, it is often impossible to find one single solution to op
timize all the tasks, since different tasks might conflict with each other. Rece
ntly, a novel method is proposed to find one single Pareto optimal solution with
 good trade-off among different tasks by casting multi-task learning as multiobj
ective optimization. In this paper, we generalize this idea and propose a novel
Pareto multi-task learning algorithm (Pareto MTL) to find a set of well-distribu
ted Pareto solutions which can represent different trade-offs among different ta
sks. The proposed algorithm first formulates a multi-task learning problem as a
multiobjective optimization problem, and then decomposes the multiobjective opti
mization problem into a set of constrained subproblems with different trade-off
preferences. By solving these subproblems in parallel, Pareto MTL can find a set
 of well-representative Pareto optimal solutions with different trade-off among
all tasks. Practitioners can easily select their preferred solution from these P
areto solutions, or use different trade-off solutions for different situations.
 Experimental results confirm that the proposed algorithm can generate well-repr
esentative solutions and outperform some state-of-the-art algorithms on many mul
ti-task learning applications.
****
Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Video
s
Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, Wenwu Zhu
Temporal sentence grounding in videos aims to detect and localize one target vid
eo segment, which semantically corresponds to a given sentence. Existing methods
 mainly tackle this task via matching and aligning semantics between a sentence
and candidate video segments, while neglect the fact that the sentence informati
on plays an important role in temporally correlating and composing the described
 contents in videos. In this paper, we propose a novel semantic conditioned dyna
mic modulation (SCDM) mechanism, which relies on the sentence semantics to modul
ate the temporal convolution operations for better correlating and composing the
 sentence related video contents over time. More importantly, the proposed SCDM
performs dynamically with respect to the diverse video contents so as to establi
sh a more precise matching relationship between sentence and video, thereby impr
oving the temporal grounding accuracy. Extensive experiments on three public dat
asets demonstrate that our proposed model outperforms the state-of-the-arts with
 clear margins, illustrating the ability of SCDM to better associate and localiz
e relevant video contents for temporal sentence grounding. Our code for this pap
er is available at https://github.com/yytzsy/SCDM.
****
Structured Variational Inference in Continuous Cox Process Models
Virginia Aglietti, Edwin V. Bonilla, Theodoros Damoulas, Sally Cripps
We propose a scalable framework for inference in a continuous sigmoidal Cox proc
ess that assumes the corresponding intensity function is given by a Gaussian pro
cess (GP) prior transformed with a scaled logistic sigmoid function. We presen
t a tractable representation of the likelihood through augmentation with a supe

rposition of Poisson processes. This view enables a structured variational approximation capturing dependencies across variables in the model. Our framework avoids discretization of the domain, does not require accurate numerical integration over the input space and is not limited to GPs with squared exponential kernels. We evaluate our approach on synthetic and real-world data showing that its benefits are particularly pronounced on multivariate input settings where it overcomes the limitations of mean-field methods and sampling schemes. We provide the state of-the-art in terms of speed, accuracy and uncertainty quantification trade-offs.

****

Channel Gating Neural Networks

Weizhe Hua, Yuan Zhou, Christopher M. De Sa, Zhiru Zhang, G. Edward Suh

This paper introduces channel gating, a dynamic, fine-grained, and hardware■efficient pruning scheme to reduce the computation cost for convolutional neural networks (CNNs). Channel gating identifies regions in the features that contribute less to the classification result, and skips the computation on a subset of the input channels for these ineffective regions. Unlike static network pruning, channel gating optimizes CNN inference at run-time by exploiting input-specific characteristics, which allows substantially reducing the compute cost with almost no accuracy loss. We experimentally show that applying channel gating in state-of-the-art networks achieves 2.7-8.0x reduction in floating-point operations (FLOPs) and 2.0-4.4x reduction in off-chip memory accesses with a minimal accuracy loss on CIFAR-10. Combining our method with knowledge distillation reduces the compute cost of ResNet-18 by 2.6x without accuracy drop on ImageNet. We further demonstrate that channel gating can be realized in hardware efficiently. Our approach exhibits sparsity patterns that are well-suited to dense systolic arrays with minimal additional hardware. We have designed an accelerator for channel gating networks, which can be implemented using either FPGAs or ASICs. Running a quantized ResNet-18 model for ImageNet, our accelerator achieves an encouraging speedup of 2.4x on average, with a theoretical FLOP reduction of 2.8x.

****

Rethinking Generative Mode Coverage: A Pointwise Guaranteed Approach

Peilin Zhong, Yuchen Mo, Chang Xiao, Pengyu Chen, Changxi Zheng

Many generative models have to combat missing modes. The conventional wisdom to this end is by reducing through training a statistical distance (such as f -divergence) between the generated distribution and provided data distribution. But this is more of a heuristic than a guarantee. The statistical distance measures a global, but not local, similarity between two distributions. Even if it is small, it does not imply a plausible mode coverage. Rethinking this problem from a game-theoretic perspective, we show that a complete mode coverage is firmly attainable. If a generative model can approximate a data distribution moderately well under a global statistical distance measure, then we will be able to find a mixture of generators that collectively covers every data point and thus every mode, with a lower-bounded generation probability. Constructing the generator mixture has a connection to the multiplicative weights update rule, upon which we propose our algorithm. We prove that our algorithm guarantees complete mode coverage. And our experiments on real and synthetic datasets confirm better mode coverage over recent approaches, ones that also use generator mixtures but rely on global statistical distances.

****

Differentially Private Algorithms for Learning Mixtures of Separated Gaussians

Gautam Kamath, Or Sheffet, Vikrant Singhal, Jonathan Ullman

Learning the parameters of Gaussian mixture models is a fundamental and widely studied problem with numerous applications. In this work, we give new algorithms for learning the parameters of a high-dimensional, well separated, Gaussian mixture model subject to the strong constraint of differential privacy. In particular, we give a differentially private analogue of the algorithm of Achlioptas and McSherry. Our algorithm has two key properties not achieved by prior work: (1) The algorithm’s sample complexity matches that of the corresponding non-private algorithm up to lower order terms in a wide range of parameters. (2) The algorith

m requires very weak a priori bounds on the parameters of the mixture components .
****
A Domain Agnostic Measure for Monitoring and Evaluating GANs
Paulina Grnarova, Kfir Y. Levy, Aurelien Lucchi, Nathanael Perraudin, Ian Goodfellow, Thomas Hofmann, Andreas Krause
Generative Adversarial Networks (GANs) have shown remarkable results in modeling complex distributions, but their evaluation remains an unsettled issue. Evaluations are essential for: (i) relative assessment of different models and (ii) monitoring the progress of a single model throughout training. The latter cannot be determined by simply inspecting the generator and discriminator loss curves as they behave non-intuitively. We leverage the notion of duality gap from game theory to propose a measure that addresses both (i) and (ii) at a low computational cost. Extensive experiments show the effectiveness of this measure to rank different GAN models and capture the typical GAN failure scenarios, including mode collapse and non-convergent behaviours. This evaluation metric also provides meaningful monitoring on the progression of the loss during training. It highly correlates with FID on natural image datasets, and with domain specific scores for text, sound and cosmology data where FID is not directly suitable. In particular, our proposed metric requires no labels or a pretrained classifier, making it domain agnostic.
****
Enabling hyperparameter optimization in sequential autoencoders for spiking neural data
Mohammad Reza Keshtkaran, Chethan Pandarinath
Continuing advances in neural interfaces have enabled simultaneous monitoring of spiking activity from hundreds to thousands of neurons. To interpret these large-scale data, several methods have been proposed to infer latent dynamic structure from high-dimensional datasets. One recent line of work uses recurrent neural networks in a sequential autoencoder (SAE) framework to uncover dynamics. SAEs are an appealing option for modeling nonlinear dynamical systems, and enable a precise link between neural activity and behavior on a single-trial basis. However, the very large parameter count and complexity of SAEs relative to other models has caused concern that SAEs may only perform well on very large training sets. We hypothesized that with a method to systematically optimize hyperparameters (HPs), SAEs might perform well even in cases of limited training data. Such a breakthrough would greatly extend their applicability. However, we find that SAEs applied to spiking neural data are prone to a particular form of overfitting that cannot be detected using standard validation metrics, which prevents standard HP searches. We develop and test two potential solutions: an alternate validation method (“sample validation”) and a novel regularization method (“coordinated dropout”). These innovations prevent overfitting quite effectively, and allow us to test whether SAEs can achieve good performance on limited data through large-scale HP optimization. When applied to data from motor cortex recorded while monkeys made reaches in various directions, large-scale HP optimization allowed SAEs to better maintain performance for small dataset sizes. Our results should greatly extend the applicability of SAEs in extracting latent dynamics from sparse, multidimensional data, such as neural population spiking activity.
****
Grid Saliency for Context Explanations of Semantic Segmentation
Lukas Hoyer, Mauricio Munoz, Prateek Katiyar, Anna Khoreva, Volker Fischer
Recently, there has been a growing interest in developing saliency methods that provide visual explanations of network predictions. Still, the usability of existing methods is limited to image classification models. To overcome this limitation, we extend the existing approaches to generate grid saliencies, which provide spatially coherent visual explanations for (pixel-level) dense prediction networks. As the proposed grid saliency allows to spatially disentangle the object and its context, we specifically explore its potential to produce context explanations for semantic segmentation networks, discovering which context most influences the class predictions inside a target object area. We investigate the effect

iveness of grid saliency on a synthetic dataset with an artificially induced bias between objects and their context as well as on the real-world Cityscapes dataset using state-of-the-art segmentation networks. Our results show that grid saliency can be successfully used to provide easily interpretable context explanations and, moreover, can be employed for detecting and localizing contextual biases present in the data.
****
Extreme Classification in Log Memory using Count-Min Sketch: A Case Study of Amazon Search with 50M Products
Tharun Kumar Reddy Medini, Qixuan Huang, Yiqiu Wang, Vijai Mohan, Anshumali Shrivastava
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.
****
Selecting the independent coordinates of manifolds with large aspect ratios
Yu-Chia Chen, Marina Meila
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.
****
DM2C: Deep Mixed-Modal Clustering
Yangbangyan Jiang, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, Qingming Huang
Data exhibited with multiple modalities are ubiquitous in real-world clustering tasks. Most existing methods, however, pose a strong assumption that the pairing information for modalities is available for all instances. In this paper, we consider a more challenging task where each instance is represented in only one modality, which we call mixed-modal data. Without any extra pairing supervision across modalities, it is difficult to find a universal semantic space for all of them. To tackle this problem, we present an adversarial learning framework for clustering with mixed-modal data. Instead of transforming all the samples into a joint modality-independent space, our framework learns the mappings across individual modal spaces by virtue of cycle-consistency. Through these mappings, we could easily unify all the samples into a single modal space and perform the clustering. Evaluations on several real-world mixed-modal datasets could demonstrate the superiority of our proposed framework.
****
An Improved Analysis of Training Over-parameterized Deep Neural Networks
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.
****
Stochastic Proximal Langevin Algorithm: Potential Splitting and Nonasymptotic Rates
Adil SALIM, Dmitry Kovalev, Peter Richtarik
We propose a new algorithm---Stochastic Proximal Langevin Algorithm (SPLA)---for sampling from a log concave distribution. Our method is a generalization of the Langevin algorithm to potentials expressed as the sum of one stochastic smooth term and multiple stochastic nonsmooth terms. In each iteration, our splitting technique only requires access to a stochastic gradient of the smooth term and a stochastic proximal operator for each of the nonsmooth terms. We establish nonasymptotic sublinear and linear convergence rates under convexity and strong convexity of the smooth term, respectively, expressed in terms of the KL divergence and Wasserstein distance. We illustrate the efficiency of our sampling technique through numerical simulations on a Bayesian learning task.
****

Contextual Bandits with Cross-Learning
Santiago Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab Mirrokni, 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.
****
Fast AutoAugment
Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, Sungwoong Kim
Data augmentation is an essential technique for improving generalization ability of deep learning models. Recently, AutoAugment \cite{cubuk2018autoaugment} has been proposed as an algorithm to automatically search for augmentation policies from a dataset and has significantly enhanced performances on many image recognition tasks. However, its search method requires thousands of GPU hours even for a relatively small dataset. In this paper, we propose an algorithm called Fast AutoAugment that finds effective augmentation policies via a more efficient search strategy based on density matching. In comparison to AutoAugment, the proposed algorithm speeds up the search time by orders of magnitude while achieves comparable performances on image recognition tasks with various models and datasets including CIFAR-10, CIFAR-100, SVHN, and ImageNet. Our code is open to the public by the official GitHub\footnote{\url{https://github.com/kakaobrain/fast-autoaugment}} of Kakao Brain.
****
A state-space model for inferring effective connectivity of latent neural dynamics from simultaneous EEG/fMRI
Tao Tu, John Paisley, Stefan Haufe, Paul Sajda
Inferring effective connectivity between spatially segregated brain regions is important for understanding human brain dynamics in health and disease. Non-invasive neuroimaging modalities, such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), are often used to make measurements and infer connectivity. However most studies do not consider integrating the two modalities even though each is an indirect measure of the latent neural dynamics and each has its own spatial and/or temporal limitations. In this study, we develop a linear state-space model to infer the effective connectivity in a distributed brain network based on simultaneously recorded EEG and fMRI data. Our method first identifies task-dependent and subject-dependent regions of interest (ROI) based on the analysis of fMRI data. Directed influences between the latent neural states at these ROIs are then modeled as a multivariate autogressive (MVAR) process driven by various exogenous inputs. The latent neural dynamics give rise to the observed scalp EEG measurements via a biophysically informed linear EEG forward model. We use a mean-field variational Bayesian approach to infer the posterior distribution of latent states and model parameters. The performance of the model was evaluated on two sets of simulations. Our results emphasize the importance of obtaining accurate spatial localization of ROIs from fMRI. Finally, we applied the model to simultaneously recorded EEG-fMRI data from 10 subjects during a Face-Car-House visual categorization task and compared the change in connectivity induced by different stimulus categories.
****
A Solvable High-Dimensional Model of GAN
Chuang Wang, Hong Hu, Yue Lu
We present a theoretical analysis of the training process for a single-layer GAN fed by high-dimensional input data. The training dynamics of the proposed model at both microscopic and macroscopic scales can be exactly analyzed in the high-dimensional limit. In particular, we prove that the macroscopic quantities measuring the quality of the training process converge to a deterministic process characterized by an ordinary differential equation (ODE), whereas the microscopic states containing all the detailed weights remain stochastic, whose dynamics can be described by a stochastic differential equation (SDE). This analysis provides a new perspective different from recent analyses in the limit of small learning

 rate, where the microscopic state is always considered deterministic, and the c
ontribution of noise is ignored. From our analysis, we show that the level of th
e background noise is essential to the convergence of the training process: sett
ing the noise level too strong leads to failure of feature recovery, whereas set
ting the noise too weak causes oscillation. Although this work focuses on a sim
ple copy model of GAN, we believe the analysis methods and insights developed he
re would prove useful in the theoretical understanding of other variants of GANs
 with more advanced training algorithms.
****
On The Classification-Distortion-Perception Tradeoff
Dong Liu, Haochen Zhang, Zhiwei Xiong
Signal degradation is ubiquitous, and computational restoration of degraded sign
al has been investigated for many years. Recently, it is reported that the capab
ility of signal restoration is fundamentally limited by the so-called perception
-distortion tradeoff, i.e. the distortion and the perceptual difference between
the restored signal and the ideal "original" signal cannot be made both minimal
simultaneously. Distortion corresponds to signal fidelity and perceptual differe
nce corresponds to perceptual naturalness, both of which are important metrics i
n practice. Besides, there is another dimension worthy of consideration--the sem
antic quality of the restored signal, i.e. the utility of the signal for recogni
tion purpose. In this paper, we extend the previous perception-distortion tradeo
ff to the case of classification-distortion-perception (CDP) tradeoff, where we
introduced the classification error rate of the restored signal in addition to d
istortion and perceptual difference. In particular, we consider the classificati
on error rate achieved on the restored signal using a predefined classifier as a
 representative metric for semantic quality. We rigorously prove the existence o
f the CDP tradeoff, i.e. the distortion, perceptual difference, and classificati
on error rate cannot be made all minimal simultaneously. We also provide both si
mulation and experimental results to showcase the CDP tradeoff. Our findings can
 be useful especially for computer vision research where some low-level vision t
asks (signal restoration) serve for high-level vision tasks (visual understandin
g). Our code and models have been published.
****
Variance Reduction for Matrix Games
Yair Carmon, Yujia Jin, Aaron Sidford, Kevin Tian
We present a randomized primal-dual algorithm that solves the problem minx maxy
y^T A x to additive error epsilon in time nnz(A) + sqrt{nnz(A) n} / epsilon, for
 matrix A with larger dimension n and nnz(A) nonzero entries. This improves the
best known exact gradient methods by a factor of sqrt{nnz(A) / n} and is faster
than fully stochastic gradient methods in the accurate and/or sparse regime epsi
lon < sqrt{n / nnz(A)$. Our results hold for x,y in the simplex (matrix games, l
inear programming) and for x in an \ell_2 ball and y in the simplex (perceptron
/ SVM, minimum enclosing ball). Our algorithm combines the Nemirovski's "concept
ual prox-method" and a novel reduced-variance gradient estimator based on "sampl
ing from the difference" between the current iterate and a reference point.
****
Efficient Forward Architecture Search
Hanzhang Hu, John Langford, Rich Caruana, Saurajit Mukherjee, Eric J. Horvitz, D
ebadeepta Dey
We propose a neural architecture search (NAS) algorithm, Petridish, to iterative
ly
add shortcut connections to existing network layers. The added shortcut connecti
ons
effectively perform gradient boosting on the augmented layers.
The proposed algorithm is motivated by the feature selection algorithm
forward stage-wise linear regression, since we consider NAS as a generalization
of feature selection for regression, where NAS selects shortcuts among layers
instead of selecting features.
In order to reduce the number of trials of possible connection combinations, we
train

jointly all possible connections at each stage of growth while leveraging
feature selection techniques to choose a subset of them.
We experimentally show this process to be an efficient forward
architecture search algorithm that can find competitive
models using few GPU days in both the search space of repeatable
network modules (cell-search) and the space of general networks (macro-search).
Petridish is particularly well-suited for warm-starting from existing models
crucial for lifelong-learning scenarios.
****
Effective End-to-end Unsupervised Outlier Detection via Inlier Priority of Discr
iminative Network
Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, Marius Klo
ft
Despite the wide success of deep neural networks (DNN), little progress has been
 made on end-to-end unsupervised outlier detection (UOD) from high dimensional d
ata like raw images. In this paper, we propose a framework named E^3Outlier, whi
ch can perform UOD in a both effective and end-to-end manner: First, instead of
the commonly-used autoencoders in previous end-to-end UOD methods, E^3Outlier fo
r the first time leverages a discriminative DNN for better representation learni
ng, by using surrogate supervision to create multiple pseudo classes from origin
al unlabelled data. Next, unlike classic UOD that utilizes data characteristics
like density or proximity, we exploit a novel property named inlier priority to
enable end-to-end UOD by discriminative DNN. We demonstrate theoretically and em
pirically that the intrinsic class imbalance of inliers/outliers will make the n
etwork prioritize minimizing inliers' loss when inliers/outliers are indiscrimin
ately fed into the network for training, which enables us to differentiate outli
ers directly from DNN's outputs. Finally, based on inlier priority, we propose t
he negative entropy based score as a simple and effective outlierness measure. E
xtensive evaluations show that E^3Outlier significantly advances UOD performance
 by up to 30% AUROC against state-of-the-art counterparts, especially on relativ
ely difficult benchmarks.
****
Poincaré Recurrence, Cycles and Spurious Equilibria in Gradient-Descent-Ascent f
or Non-Convex Non-Concave Zero-Sum Games
Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Georgios Piliouras
We study a wide class of non-convex non-concave min-max games that generalizes o
ver standard bilinear zero-sum games. In this class, players control the inputs
of a smooth function whose output is being applied to a bilinear zero-sum game.
This class of games is motivated by the indirect nature of the competition in
Generative Adversarial Networks, where players control the parameters of a neura
l network while the actual competition happens between the distributions that th
e generator and discriminator capture. We establish theoretically, that dependin
g on the specific instance of the problem gradient-descent-ascent dynamics can
exhibit a variety of behaviors antithetical to convergence to the game theoretic
ally meaningful min-max solution. Specifically, different forms of recurrent beh
avior (including periodicity and Poincar\'{e} recurrence) are possible as well a
s convergence to spurious (non-min-max) equilibria for a positive measure of ini
tial conditions. At the technical level, our analysis combines tools from optimi
zation theory, game theory and dynamical systems.
****
End-to-End Learning on 3D Protein Structure for Interface Prediction
Raphael Townshend, Rishi Bedi, Patricia Suriana, Ron Dror
Despite an explosion in the number of experimentally determined, atomically deta
iled structures of biomolecules, many critical tasks in structural biology remai
n data-limited. Whether performance in such tasks can be improved by using larg
e repositories of tangentially related structural data remains an open question.
  To address this question, we focused on a central problem in biology: predicti
ng how proteins interact with one another—that is, which surfaces of one protein
 bind to those of another protein. We built a training dataset, the Database of
 Interacting Protein Structures (DIPS), that contains biases but is two orders o

f magnitude larger than those used previously. We found that these biases significantly degrade the performance of existing methods on gold-standard data. Hypothesizing that assumptions baked into the hand-crafted features on which these methods depend were the source of the problem, we developed the first end-to-end learning model for protein interface prediction, the Siamese Atomic Surfacelet Network (SASNet). Using only spatial coordinates and identities of atoms, SASNet outperforms state-of-the-art methods trained on gold-standard structural data, even when trained on only 3% of our new dataset. Code and data available at https://github.com/drorlab/DIPS.
****
Scalable Global Optimization via Local Bayesian Optimization
David Eriksson, Michael Pearce, Jacob Gardner, Ryan D. Turner, Matthias Poloczek
Bayesian optimization has recently emerged as a popular method for the sample-efficient optimization of expensive black-box functions. However, the application to high-dimensional problems with several thousand observations remains challenging, and on difficult problems Bayesian optimization is often not competitive with other paradigms. In this paper we take the view that this is due to the implicit homogeneity of the global probabilistic models and an overemphasized exploration that results from global acquisition. This motivates the design of a local probabilistic approach for global optimization of large-scale high-dimensional problems. We propose the TuRBO algorithm that fits a collection of local models and performs a principled global allocation of samples across these models via an implicit bandit approach. A comprehensive evaluation demonstrates that TuRBO outperforms state-of-the-art methods from machine learning and operations research on problems spanning reinforcement learning, robotics, and the natural sciences.
****
Positional Normalization
Boyi Li, Felix Wu, Kilian Q. Weinberger, Serge Belongie
A widely deployed method for reducing the training time of deep neural networks is to normalize activations at each layer. Although various normalization schemes have been proposed, they all follow a common theme: normalize across spatial dimensions and discard the extracted statistics. In this paper, we propose a novel normalization method that deviates from this theme. Our approach, which we refer to as Positional Normalization (PONO), normalizes exclusively across channels, which allows us to capture structural information of the input image in the first and second moments. Instead of disregarding this information, we inject it into later layers to preserve or transfer structural information in generative networks. We show that PONO significantly improves the performance of deep networks across a wide range of model architectures and image generation tasks.
****
Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model
Atilim Gunes Baydin, Lei Shao, Wahid Bhimji, Lukas Heinrich, Saeid Naderiparizi, Andreas Munk, Jialin Liu, Bradley Gram-Hansen, Gilles Louppe, Lawrence Meadows, Philip Torr, Victor Lee, Kyle Cranmer, Mr. Prabhat, Frank Wood
We present a novel probabilistic programming framework that couples directly to existing large-scale simulators through a cross-platform probabilistic execution protocol, which allows general-purpose inference engines to record and control random number draws within simulators in a language-agnostic way. The execution of existing simulators as probabilistic programs enables highly interpretable posterior inference in the structured model defined by the simulator code base. We demonstrate the technique in particle physics, on a scientifically accurate simulation of the tau lepton decay, which is a key ingredient in establishing the properties of the Higgs boson. Inference efficiency is achieved via inference compilation where a deep recurrent neural network is trained to parameterize proposal distributions and control the stochastic simulator in a sequential importance sampling scheme, at a fraction of the computational cost of a Markov chain Monte Carlo baseline.
****

Online Optimal Control with Linear Dynamics and Predictions: Algorithms and Regret Analysis
Yingying Li, Xin Chen, Na Li
This paper studies the online optimal control problem with time-varying convex stage costs for a time-invariant linear dynamical system, where a finite lookahead window of accurate predictions of the stage costs are available at each time. We design online algorithms, Receding Horizon Gradient-based Control (RHGC), that utilize the predictions through finite steps of gradient computations. We study the algorithm performance measured by dynamic regret: the online performance minus the optimal performance in hindsight. It is shown that the dynamic regret of RHGC decays exponentially with the size of the lookahead window. In addition, we provide a fundamental limit of the dynamic regret for any online algorithms by considering linear quadratic tracking problems. The regret upper bound of one RHGC method almost reaches the fundamental limit, demonstrating the effectiveness of the algorithm. Finally, we numerically test our algorithms for both linear and nonlinear systems to show the effectiveness and generality of our RHGC.
****
Beyond Vector Spaces: Compact Data Representation as Differentiable Weighted Graphs
Denis Mazur, Vage Egiazarian, Stanislav Morozov, Artem Babenko
Learning useful representations is a key ingredient to the success of modern machine learning. Currently, representation learning mostly relies on embedding data into Euclidean space. However, recent work has shown that data in some domains is better modeled by non-euclidean metric spaces, and inappropriate geometry can result in inferior performance. In this paper, we aim to eliminate the inductive bias imposed by the embedding space geometry. Namely, we propose to map data into more general non-vector metric spaces: a weighted graph with a shortest path distance. By design, such graphs can model arbitrary geometry with a proper configuration of edges and weights. Our main contribution is PRODIGE: a method that learns a weighted graph representation of data end-to-end by gradient descent. Greater generality and fewer model assumptions make PRODIGE more powerful than existing embedding-based approaches. We confirm the superiority of our method via extensive experiments on a wide range of tasks, including classification, compression, and collaborative filtering.
****
Gradient Information for Representation and Modeling
Jie Ding, Robert Calderbank, Vahid Tarokh
Motivated by Fisher divergence, in this paper we present a new set of information quantities which we refer to as gradient information. These measures serve as surrogates for classical information measures such as those based on logarithmic loss, Kullback-Leibler divergence, directed Shannon information, etc. in many data-processing scenarios of interest, and often provide significant computational advantage, improved stability and robustness. As an example, we apply these measures to the Chow-Liu tree algorithm, and demonstrate remarkable performance and significant computational reduction using both synthetic and real data.
****
Category Anchor-Guided Unsupervised Domain Adaptation for Semantic Segmentation
Qiming ZHANG, Jing Zhang, Wei Liu, Dacheng Tao
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.
****
Novel positional encodings to enable tree-based transformers
Vighnesh Shiv, Chris Quirk
Neural models optimized for tree-based problems are of great value in tasks like SQL query extraction and program synthesis.
On sequence-structured data, transformers have been shown to learn relationships across arbitrary pairs of positions more reliably than recurrent models.
Motivated by this property, we propose a method to extend transformers to tree-s

tructured data, enabling sequence-to-tree, tree-to-sequence, and tree-to-tree mappings.
Our approach abstracts the transformer's sinusoidal positional encodings, allowing us to instead use a novel positional encoding scheme to represent node positions within trees.
We evaluated our model in tree-to-tree program translation and sequence-to-tree semantic parsing settings, achieving superior performance over both sequence-to-sequence transformers and state-of-the-art tree-based LSTMs on several datasets. In particular, our results include a 22% absolute increase in accuracy on a JavaScript to CoffeeScript translation dataset.
****
Algorithm-Dependent Generalization Bounds for Overparameterized Deep Residual Networks
Spencer Frei, Yuan Cao, Quanquan Gu
The skip-connections used in residual networks have become a standard architecture choice in deep learning due to the increased generalization and stability of networks with this architecture, although there have been limited theoretical guarantees for this improved performance. In this work, we analyze overparameterized deep residual networks trained by gradient descent following random initialization, and demonstrate that (i) the class of networks learned by gradient descent constitutes a small subset of the entire neural network function class, and (ii) this subclass of networks is sufficiently large to guarantee small training error. By showing (i) we are able to demonstrate that deep residual networks trained with gradient descent have a small generalization gap between training and test error, and together with (ii) this guarantees that the test error will be small. Our optimization and generalization guarantees require overparameterization that is only logarithmic in the depth of the network, which helps explain why residual networks are preferable to fully connected ones.
****
Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching
Hongteng Xu, Dixin Luo, Lawrence Carin
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.
****
Riemannian batch normalization for SPD neural networks
Daniel Brooks, Olivier Schwander, Frederic Barbaresco, Jean-Yves Schneider, Matthieu Cord
Covariance matrices have attracted attention for machine learning applications due
to their capacity to capture interesting structure in the data. The main challenge
is that one needs to take into account the particular geometry of the Riemannian manifold of symmetric positive definite (SPD) matrices they belong to. In the con-
text of deep networks, several architectures for these matrices have recently been
proposed. In our article, we introduce a Riemannian batch normalization (batch-norm) algorithm, which generalizes the one used in Euclidean nets. This novel layer makes use of geometric operations on the manifold, notably the Riemannian barycenter, parallel transport and non-linear structured matrix transformations. We
derive a new manifold-constrained gradient descent algorithm working in the space
of SPD matrices, allowing to learn the batchnorm layer. We validate our proposed approach with experiments in three different contexts on diverse data types: a drone recognition dataset from radar observations, and on emotion and action recognition datasets from video and motion capture data. Experiments show that the Riemannian batchnorm systematically gives better classification performance

compared with leading methods and a remarkable robustness to lack of data.
****
Deep Set Prediction Networks
Yan Zhang, Jonathon Hare, Adam Prugel-Bennett
Current approaches for predicting sets from feature vectors ignore the unordered nature of sets and suffer from discontinuity issues as a result. We propose a general model for predicting sets that properly respects the structure of sets and avoids this problem. With a single feature vector as input, we show that our model is able to auto-encode point sets, predict the set of bounding boxes of objects in an image, and predict the set of attributes of these objects.
****
A unified theory for the origin of grid cells through the lens of pattern formation
Ben Sorscher, Gabriel Mel, Surya Ganguli, Samuel Ocko
Grid cells in the brain fire in strikingly regular hexagonal patterns across space. There are currently two seemingly unrelated frameworks for understanding these patterns. Mechanistic models account for hexagonal firing fields as the result of pattern-forming dynamics in a recurrent neural network with hand-tuned center-surround connectivity. Normative models specify a neural architecture, a learning rule, and a navigational task, and observe that grid-like firing fields emerge due to the constraints of solving this task. Here we provide an analytic theory that unifies the two perspectives by casting the learning dynamics of neural networks trained on navigational tasks as a pattern forming dynamical system. This theory provides insight into the optimal solutions of diverse formulations of the normative task, and shows that symmetries in the representation of space correctly predict the structure of learned firing fields in trained neural networks. Further, our theory proves that a nonnegativity constraint on firing rates induces a symmetry-breaking mechanism which favors hexagonal firing fields. We extend this theory to the case of learning multiple grid maps and demonstrate that optimal solutions consist of a hierarchy of maps with increasing length scales. These results unify previous accounts of grid cell firing and provide a novel framework for predicting the learned representations of recurrent neural networks.
****
Functional Adversarial Attacks
Cassidy Laidlaw, Soheil Feizi
We propose functional adversarial attacks, a novel class of threat models for crafting adversarial examples to fool machine learning models. Unlike a standard lp-ball threat model, a functional adversarial threat model allows only a single function to be used to perturb input features to produce an adversarial example. For example, a functional adversarial attack applied on colors of an image can change all red pixels simultaneously to light red. Such global uniform changes in images can be less perceptible than perturbing pixels of the image individually. For simplicity, we refer to functional adversarial attacks on image colors as ReColorAdv, which is the main focus of our experiments. We show that functional threat models can be combined with existing additive (lp) threat models to generate stronger threat models that allow both small, individual perturbations and large, uniform changes to an input. Moreover, we prove that such combinations encompass perturbations that would not be allowed in either constituent threat model. In practice, ReColorAdv can significantly reduce the accuracy of a ResNet-32 trained on CIFAR-10. Furthermore, to the best of our knowledge, combining ReColorAdv with other attacks leads to the strongest existing attack even after adversarial training.
****
Memory-oriented Decoder for Light Field Salient Object Detection
Miao Zhang, Jingjing Li, JI WEI, Yongri Piao, Huchuan Lu
Light field data have been demonstrated in favor of many tasks in computer vision, but existing works about light field saliency detection still rely on hand-crafted features. In this paper, we present a deep-learning-based method where a novel memory-oriented decoder is tailored for light field saliency detection. Our

goal is to deeply explore and comprehensively exploit internal correlation of focal slices for accurate prediction by designing feature fusion and integration mechanisms. The success of our method is demonstrated by achieving the state of the art on three datasets. We present this problem in a way that is accessible to members of the community and provide a large-scale light field dataset that facilitates comparisons across algorithms. The code and dataset will be made publicly available.
****
Learning search spaces for Bayesian optimization: Another view of hyperparameter transfer learning
Valerio Perrone, Huibin Shen, Matthias W. Seeger, Cedric Archambeau, Rodolphe Jenatton
Bayesian optimization (BO) is a successful methodology to optimize black-box functions that are expensive to evaluate. While traditional methods optimize each black-box function in isolation, there has been recent interest in speeding up BO by transferring knowledge across multiple related black-box functions. In this work, we introduce a method to automatically design the BO search space by relying on evaluations of previous black-box functions. We depart from the common practice of defining a set of arbitrary search ranges a priori by considering search space geometries that are learnt from historical data. This simple, yet effective strategy can be used to endow many existing BO methods with transfer learning properties. Despite its simplicity, we show that our approach considerably boosts BO by reducing the size of the search space, thus accelerating the optimization of a variety of black-box optimization problems. In particular, the proposed approach combined with random search results in a parameter-free, easy-to-implement, robust hyperparameter optimization strategy. We hope it will constitute a natural baseline for further research attempting to warm-start BO.
****
Image Synthesis with a Single (Robust) Classifier
Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry
We show that the basic classification framework alone can be used to tackle some of the most challenging tasks in image synthesis. In contrast to other state-of-the-art approaches, the toolkit we develop is rather minimal: it uses a single, off-the-shelf classifier for all these tasks. The crux of our approach is that we train this classifier to be adversarially robust. It turns out that adversarial robustness is precisely what we need to directly manipulate salient features of the input. Overall, our findings demonstrate the utility of robustness in the broader machine learning context.
****
Learning from Trajectories via Subgoal Discovery
Sujoy Paul, Jeroen Vanbaar, Amit Roy-Chowdhury
Learning to solve complex goal-oriented tasks with sparse terminal-only rewards often requires an enormous number of samples. In such cases, using a set of expert trajectories could help to learn faster. However, Imitation Learning (IL) via supervised pre-training with these trajectories may not perform as well and generally requires additional finetuning with expert-in-the-loop. In this paper, we propose an approach which uses the expert trajectories and learns to decompose the complex main task into smaller sub-goals. We learn a function which partitions the state-space into sub-goals, which can then be used to design an extrinsic reward function. We follow a strategy where the agent first learns from the trajectories using IL and then switches to Reinforcement Learning (RL) using the identified sub-goals, to alleviate the errors in the IL step. To deal with states which are under-represented by the trajectory set, we also learn a function to modulate the sub-goal predictions. We show that our method is able to solve complex goal-oriented tasks, which other RL, IL or their combinations in literature are not able to solve.
****
Unsupervised State Representation Learning in Atari
Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, R D

evon Hjelm
State representation learning, or the ability to capture latent generative factors of an environment is crucial for building intelligent agents that can perform a wide variety of tasks. Learning such representations in an unsupervised manner without supervision from rewards is an open problem. We introduce a method that tries to learn better state representations by maximizing mutual information across spatially and temporally distinct features of a neural encoder of the observations. We also introduce a new benchmark based on Atari 2600 games where we evaluate representations based on how well they capture the ground truth state. We believe this new framework for evaluating representation learning models will be crucial for future representation learning research. Finally, we compare our technique with other state-of-the-art generative and contrastive representation learning methods.
****
Loaded DiCE: Trading off Bias and Variance in Any-Order Score Function Gradient Estimators for Reinforcement Learning
Gregory Farquhar, Shimon Whiteson, Jakob Foerster
Gradient-based methods for optimisation of objectives in stochastic settings with unknown or intractable dynamics require estimators of derivatives. We derive an objective that, under automatic differentiation, produces low-variance unbiased estimators of derivatives at any order. Our objective is compatible with arbitrary advantage estimators, which allows the control of the bias and variance of any-order derivatives when using function approximation. Furthermore, we propose a method to trade off bias and variance of higher order derivatives by discounting the impact of more distant causal dependencies. We demonstrate the correctness and utility of our estimator in analytically tractable MDPs and in meta-reinforcement-learning for continuous control.
****
Meta Learning with Relational Information for Short Sequences
Yujia Xie, Haoming Jiang, Feng Liu, Tuo Zhao, Hongyuan Zha
This paper proposes a new meta-learning method -- named HARMLESS (HAwkes Relational Meta Learning method for Short Sequences) for learning heterogeneous point process models from a collection of short event sequence data along with a relational network. Specifically, we propose a hierarchical Bayesian mixture Hawkes process model, which naturally incorporates the relational information among sequences into point process modeling. Compared with existing methods, our model can capture the underlying mixed-community patterns of the relational network, which simultaneously encourages knowledge sharing among sequences and facilitates adaptively learning for each individual sequence. We further propose an efficient stochastic variational meta-EM algorithm, which can scale to large problems. Numerical experiments on both synthetic and real data show that HARMLESS outperforms existing methods in terms of predicting the future events.
****
Private Learning Implies Online Learning: An Efficient Reduction
Alon Gonen, Elad Hazan, Shay Moran
We study the relationship between the notions of differentially private learning and online learning. Several recent works have shown that differentially private learning implies online learning, but an open problem of Neel, Roth, and Wu \cite{NeelAaronRoth2018} asks whether this implication is {\it efficient}.
Specifically, does an efficient differentially private learner imply an efficient online learner?
****
Learning from brains how to regularize machines
Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, Andreas Tolias
Despite impressive performance on numerous visual tasks, Convolutional Neural Networks (CNNs) --- unlike brains --- are often highly sensitive to small perturbations of their input, e.g. adversarial noise leading to erroneous decisions. We propose to regularize CNNs using large-scale neuroscience data to learn more robust neural features in terms of representational similarity. We presented natura

l images to mice and measured the responses of thousands of neurons from cortical visual areas. Next, we denoised the notoriously variable neural activity using strong predictive models trained on this large corpus of responses from the mouse visual system, and calculated the representational similarity for millions of pairs of images from the model's predictions. We then used the neural representation similarity to regularize CNNs trained on image classification by penalizing intermediate representations that deviated from neural ones. This preserved performance of baseline models when classifying images under standard benchmarks, while maintaining substantially higher performance compared to baseline or control models when classifying noisy images. Moreover, the models regularized with cortical representations also improved model robustness in terms of adversarial attacks. This demonstrates that regularizing with neural data can be an effective tool to create an inductive bias towards more robust inference.
****

Kernel quadrature with DPPs
Ayoub Belhadji, Rémi Bardenet, Pierre Chainais
We study quadrature rules for functions living in an RKHS, using nodes sampled from a projection determinantal point process (DPP). DPPs are parametrized by a kernel, and we use a truncated and saturated version of the RKHS kernel.
This natural link between the two kernels, along with DPP machinery, leads to relatively tight bounds on the quadrature error, that depends on the spectrum of the RKHS kernel. Finally, we experimentally compare DPPs to existing kernel-based quadratures such as herding, Bayesian quadrature, or continuous leverage score sampling. Numerical results confirm the interest of DPPs, and even suggest faster rates than our bounds in particular cases.
****

A Debiased MDI Feature Importance Measure for Random Forests
Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, Bin Yu
Tree ensembles such as Random Forests have achieved impressive empirical success across a wide variety of applications. To understand how these models make predictions, people routinely turn to feature importance measures calculated from tree ensembles. It has long been known that Mean Decrease Impurity (MDI), one of the most widely used measures of feature importance, incorrectly assigns high importance to noisy features, leading to systematic bias in feature selection. In this paper, we address the feature selection bias of MDI from both theoretical and methodological perspectives. Based on the original definition of MDI by Breiman et al. \cite{Breiman1984} for a single tree, we derive a tight non-asymptotic bound on the expected bias of MDI importance of noisy features, showing that deep trees have higher (expected) feature selection bias than shallow ones. However, it is not clear how to reduce the bias of MDI using its existing analytical expression. We derive a new analytical expression for MDI, and based on this new expression, we are able to propose a debiased MDI feature importance measure using out-of-bag samples, called MDI-oob. For both the simulated data and a genomic ChIP dataset, MDI-oob achieves state-of-the-art performance in feature selection from Random Forests for both deep and shallow trees.
****

Unsupervised Discovery of Temporal Structure in Noisy Data with Dynamical Components Analysis
David Clark, Jesse Livezey, Kristofer Bouchard
Linear dimensionality reduction methods are commonly used to extract low-dimensional structure from high-dimensional data. However, popular methods disregard temporal structure, rendering them prone to extracting noise rather than meaningful dynamics when applied to time series data. At the same time, many successful unsupervised learning methods for temporal, sequential and spatial data extract features which are predictive of their surrounding context. Combining these approaches, we introduce Dynamical Components Analysis (DCA), a linear dimensionality reduction method which discovers a subspace of high-dimensional time series data with maximal predictive information, defined as the mutual information between the past and future. We test DCA on synthetic examples and demonstrate its superior ability to extract dynamical structure compared to commonly used linear met

hods. We also apply DCA to several real-world datasets, showing that the dimensions extracted by DCA are more useful than those extracted by other methods for predicting future states and decoding auxiliary variables. Overall, DCA robustly extracts dynamical structure in noisy, high-dimensional data while retaining the computational efficiency and geometric interpretability of linear dimensionality reduction methods.
****
MintNet: Building Invertible Neural Networks with Masked Convolutions
Yang Song, Chenlin Meng, Stefano Ermon
We propose a new way of constructing invertible neural networks by combining simple building blocks with a novel set of composition rules. This leads to a rich set of invertible architectures, including those similar to ResNets. Inversion is achieved with a locally convergent iterative procedure that is parallelizable and very fast in practice. Additionally, the determinant of the Jacobian can be computed analytically and efficiently, enabling their generative use as flow models. To demonstrate their flexibility, we show that our invertible neural networks are competitive with ResNets on MNIST and CIFAR-10 classification. When trained as generative models, our invertible networks achieve competitive likelihoods on MNIST, CIFAR-10 and ImageNet 32x32, with bits per dimension of 0.98, 3.32 and 4.06 respectively.
****
Learning Temporal Pose Estimation from Sparsely-Labeled Videos
Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, Lorenzo Torresani
Modern approaches for multi-person pose estimation in video require large amounts of dense annotations. However, labeling every frame in a video is costly and labor intensive. To reduce the need for dense annotations, we propose a PoseWarper network that leverages training videos with sparse annotations (every k frames) to learn to perform dense temporal pose propagation and estimation. Given a pair of video frames---a labeled Frame A and an unlabeled Frame B---we train our model to predict human pose in Frame A using the features from Frame B by means of deformable convolutions to implicitly learn the pose warping between A and B. We demonstrate that we can leverage our trained PoseWarper for several applications. First, at inference time we can reverse the application direction of our network in order to propagate pose information from manually annotated frames to unlabeled frames. This makes it possible to generate pose annotations for the entire video given only a few manually-labeled frames. Compared to modern label propagation methods based on optical flow, our warping mechanism is much more compact (6M vs 39M parameters), and also more accurate (88.7% mAP vs 83.8% mAP). We also show that we can improve the accuracy of a pose estimator by training it on an augmented dataset obtained by adding our propagated poses to the original manual labels. Lastly, we can use our PoseWarper to aggregate temporal pose information from neighboring frames during inference. This allows us to obtain state-of-the-art pose detection results on PoseTrack2017 and PoseTrack2018 datasets.
****
Learning Generalizable Device Placement Algorithms for Distributed Machine Learning
ravichandra addanki, Shaileshh Bojja Venkatakrishnan, Shreyan Gupta, Hongzi Mao, Mohammad Alizadeh
We present Placeto, a reinforcement learning (RL) approach to
efficiently find device placements for distributed neural network training.
****
Dynamic Incentive-Aware Learning: Robust Pricing in Contextual Auctions
Negin Golrezaei, Adel Javanmard, Vahab Mirrokni
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.
****
Optimal Best Markovian Arm Identification with Fixed Confidence
Vrettos Moulos

We give a complete characterization of the sampling complexity of best Markovian arm identification in one-parameter Markovian bandit models. We derive instance specific nonasymptotic and asymptotic lower bounds which generalize those of the IID setting.
We analyze the Track-and-Stop strategy, initially proposed for the IID setting, and we prove that asymptotically it is at most a factor of four apart from the lower bound. Our one-parameter Markovian bandit model is based on the notion of an exponential family of stochastic matrices for which we establish many useful properties. For the analysis of the Track-and-Stop strategy we derive a novel and optimal concentration inequality for Markov chains that may be of interest in its own right.
****
On the equivalence between graph isomorphism testing and function approximation with GNNs
Zhengdao Chen, Soledad Villar, Lei Chen, Joan Bruna
Graph neural networks (GNNs) have achieved lots of success on graph-structured data. In light of this, there has been increasing interest in studying their representation power. One line of work focuses on the universal approximation of permutation-invariant functions by certain classes of GNNs, and another demonstrates the limitation of GNNs via graph isomorphism tests.
****
Information Competing Process for Learning Diversified Representations
Jie Hu, Rongrong Ji, ShengChuan Zhang, Xiaoshuai Sun, Qixiang Ye, Chia-Wen Lin, Qi Tian
Learning representations with diversified information remains as an open problem. Towards learning diversified representations, a new approach, termed Information Competing Process (ICP), is proposed in this paper. Aiming to enrich the information carried by feature representations, ICP separates a representation into two parts with different mutual information constraints. The separated parts are forced to accomplish the downstream task independently in a competitive environment which prevents the two parts from learning what each other learned for the downstream task. Such competing parts are then combined synergistically to complete the task. By fusing representation parts learned competitively under different conditions, ICP facilitates obtaining diversified representations which contain rich information. Experiments on image classification and image reconstruction tasks demonstrate the great potential of ICP to learn discriminative and disentangled representations in both supervised and self-supervised learning settings.
****
Individual Regret in Cooperative Nonstochastic Multi-Armed Bandits
Yogev Bar-On, Yishay Mansour
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.
****
SPoC: Search-based Pseudocode to Code
Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, Percy S. Liang
We consider the task of mapping pseudocode to executable code, assuming a one-to-one correspondence between lines of pseudocode and lines of code. Given test cases as a mechanism to validate programs, we search over the space of possible translations of the pseudocode to find a program that compiles and passes the test cases. While performing a best-first search, compilation errors constitute 88.7% of program failures. To better guide this search, we learn to predict the line of the program responsible for the failure and focus search over alternative translations of the pseudocode for that line. For evaluation, we collected the SPoC dataset (Search-based Pseudocode to Code) containing 18,356 C++ programs with human-authored pseudocode and test cases. Under a budget of 100 program compilations, performing search improves the synthesis success rate over using the top-

one translation of the pseudocode from 25.6% to 44.7%.
****
Distributional Policy Optimization: An Alternative Approach for Continuous Contr
ol
Chen Tessler, Guy Tennenholtz, Shie Mannor
We identify a fundamental problem in policy gradient-based methods in continuous
 control. As policy gradient methods require the agent's underlying probability
distribution, they limit policy representation to parametric distribution classe
s. We show that optimizing over such sets results in local movement in the actio
n space and thus convergence to sub-optimal solutions. We suggest a novel distri
butional framework, able to represent arbitrary distribution functions over the
continuous action space. Using this framework, we construct a generative scheme,
 trained using an off-policy actor-critic paradigm, which we call the Generative
 Actor Critic (GAC). Compared to policy gradient methods, GAC does not require k
nowledge of the underlying probability distribution, thereby overcoming these li
mitations. Empirical evaluation shows that our approach is comparable and often
surpasses current state-of-the-art baselines in continuous domains.
****
Oblivious Sampling Algorithms for Private Data Analysis
Sajin Sasy, Olga Ohrimenko
We study secure and privacy-preserving data analysis
based on queries executed on samples from a dataset.
Trusted execution environments (TEEs) can be used to
protect the content of the data during query computation,
while supporting differential-private (DP) queries in TEEs
provides record privacy when query output is revealed.
Support for sample-based queries is attractive
due to \emph{privacy amplification}
since not all dataset is used to answer a query but only a small subset.
However, extracting data samples with TEEs
while proving strong DP guarantees is not
trivial as secrecy of sample indices has to be preserved.
To this end, we design efficient secure variants of common sampling algorithms.
Experimentally we show that accuracy of models
trained with shuffling and sampling is the same for
differentially private models for MNIST and CIFAR-10,
while sampling provides stronger privacy guarantees than shuffling.
****
On Relating Explanations and Adversarial Examples
Alexey Ignatiev, Nina Narodytska, Joao Marques-Silva
The importance of explanations (XP's) of machine learning (ML) model predictions
 and of adversarial examples (AE's) cannot be overstated, with both arguably bei
ng essential for the practical success of ML in different settings. There has be
en recent work on understanding and assessing the relationship between XP's and
AE's. However, such work has been mostly experimental and a sound theoretical re
lationship has been elusive. This paper demonstrates that explanations and adver
sarial examples are related by a generalized form of hitting set duality, which
extends earlier work on hitting set duality observed in model-based diagnosis an
d knowledge compilation. Furthermore, the paper proposes algorithms, which enabl
e computing adversarial examples from explanations and vice-versa.
****
Greedy Sampling for Approximate Clustering in the Presence of Outliers
Aditya Bhaskara, Sharvaree Vadgama, Hong Xu
Greedy algorithms such as adaptive sampling (k-means++) and furthest point trave
rsal are popular choices for clustering problems. One the one hand, they possess
 good theoretical approximation guarantees, and on the other, they are fast and
easy to implement. However, one main issue with these algorithms is the sensitiv
ity to noise/outliers in the data. In this work we show that for k-means and k-c
enter clustering, simple modifications to the well-studied greedy algorithms res
ult in nearly identical guarantees, while additionally being robust to outliers.

For instance, in the case of k-means++, we show that a simple thresholding operation on the distances suffices to obtain an O(\log k) approximation to the objective. We obtain similar results for the simpler k-center problem. Finally, we show experimentally that our algorithms are easy to implement and scale well. We also measure their ability to identify noisy points added to a dataset.
****
Understanding the Representation Power of Graph Neural Networks in Learning Graph Topology
Nima Dehmamy, Albert-Laszlo Barabasi, Rose Yu
To deepen our understanding of graph neural networks, we investigate the representation power of Graph Convolutional Networks (GCN) through the looking glass of graph moments, a key property of graph topology encoding path of various lengths.
We find that GCNs are rather restrictive in learning graph moments. Without careful design, GCNs can fail miserably even with multiple layers and nonlinear activation functions.
We analyze theoretically the expressiveness of GCNs, arriving at a modular GCN design, using different propagation rules.
Our modular design is capable of distinguishing graphs from different graph generation models for surprisingly small graphs, a notoriously difficult problem in network science.
Our investigation suggests that, depth is much more influential than width and deeper GCNs are more capable of learning higher order graph moments.
Additionally, combining GCN modules with different propagation rules is critical to the representation power of GCNs.
****
Single-Model Uncertainties for Deep Learning
Natasa Tagasovska, David Lopez-Paz
We provide single-model estimates of aleatoric and epistemic uncertainty for deep neural networks.
To estimate aleatoric uncertainty, we propose Simultaneous Quantile Regression (SQR), a loss function to learn all the conditional quantiles of a given target variable.
These quantiles can be used to compute well-calibrated prediction intervals.
To estimate epistemic uncertainty, we propose Orthonormal Certificates (OCs), a collection of diverse non-constant functions that map all training samples to zero.
These certificates map out-of-distribution examples to non-zero values, signaling epistemic uncertainty.
Our uncertainty estimators are computationally attractive, as they do not require ensembling or retraining deep models, and achieve state-of-the-art performance.
****
The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the XAUC Metric
Nathan Kallus, Angela Zhou
Where machine-learned predictive risk scores inform high-stakes decisions, such as bail and sentencing in criminal justice, fairness has been a serious concern. Recent work has characterized the disparate impact that such risk scores can have when used for a binary classification task. This may not account, however, for the more diverse downstream uses of risk scores and their non-binary nature. To better account for this, in this paper, we investigate the fairness of predictive risk scores from the point of view of a bipartite ranking task, where one seeks to rank positive examples higher than negative ones. We introduce the xAUC disparity as a metric to assess the disparate impact of risk scores and define it as the difference in the probabilities of ranking a random positive example from one protected group above a negative one from another group and vice versa. We provide a decomposition of bipartite ranking loss into components that involve the discrepancy and components that involve pure predictive ability within each group. We use xAUC analysis to audit predictive risk scores for recidivism predi

ction, income prediction, and cardiac arrest prediction, where it describes disparities that are not evident from simply comparing within-group predictive performance.
****
Robust Principal Component Analysis with Adaptive Neighbors
Rui Zhang, Hanghang Tong
Suppose certain data points are overly contaminated, then the existing principal component analysis (PCA) methods are frequently incapable of filtering out and eliminating the excessively polluted ones, which potentially lead to the functional degeneration of the corresponding models. To tackle the issue, we propose a general framework namely robust weight learning with adaptive neighbors (RWL-AN), via which adaptive weight vector is automatically obtained with both robustness and sparse neighbors. More significantly, the degree of the sparsity is steerable such that only exact k well-fitting samples with least reconstruction errors are activated during the optimization, while the residual samples, i.e., the extreme noised ones are eliminated for the global robustness. Additionally, the framework is further applied to PCA problem to demonstrate the superiority and effectiveness of the proposed RWL-AN model.
****
Wasserstein Weisfeiler-Lehman Graph Kernels
Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, Karsten Borgwardt
Most graph kernels are an instance of the class of R-Convolution kernels, which measure the similarity of objects by comparing their substructures.
Despite their empirical success, most graph kernels use a naive aggregation of the final set of substructures, usually a sum or average, thereby potentially discarding valuable information about the distribution of individual components. Furthermore, only a limited instance of these approaches can be extended to continuously attributed graphs.
We propose a novel method that relies on the Wasserstein distance between the node feature vector distributions of two graphs, which allows to find subtler differences in data sets by considering graphs as high-dimensional objects, rather than simple means.
We further propose a Weisfeiler--Lehman inspired embedding scheme for graphs with continuous node attributes and weighted edges, enhance it with the computed Wasserstein distance, and thus improve the state-of-the-art prediction performance on several graph classification tasks.
****
DATA: Differentiable ArchiTecture Approximation
Jianlong Chang, xinbang zhang, Yiwen Guo, GAOFENG MENG, SHIMING XIANG, Chunhong Pan
Neural architecture search (NAS) is inherently subject to the gap of architectures during searching and validating. To bridge this gap, we develop Differentiable ArchiTecture Approximation (DATA) with an Ensemble Gumbel-Softmax (EGS) estimator to automatically approximate architectures during searching and validating in a differentiable manner. Technically, the EGS estimator consists of a group of Gumbel-Softmax estimators, which is capable of converting probability vectors to binary codes and passing gradients from binary codes to probability vectors. Benefiting from such modeling, in searching, architecture parameters and network weights in the NAS model can be jointly optimized with the standard back-propagation, yielding an end-to-end learning mechanism for searching deep models in a large enough search space. Conclusively, during validating, a high-performance architecture that approaches to the learned one during searching is readily built. Extensive experiments on a variety of popular datasets strongly evidence that our method is capable of discovering high-performance architectures for image classification, language modeling and semantic segmentation, while guaranteeing the requisite efficiency during searching.
****
Near Neighbor: Who is the Fairest of Them All?
Sariel Har-Peled, Sepideh Mahabadi

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 Co-Learning on G-Manifolds Across Irreducible Representations
Yifeng Fan, Tingran Gao, Zhizhen Jane Zhao
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.
****

Fast Efficient Hyperparameter Tuning for Policy Gradient Methods
Supratik Paul, Vitaly Kurin, Shimon Whiteson
The performance of policy gradient methods is sensitive to hyperparameter settings that must be tuned for any new application. Widely used grid search methods for tuning hyperparameters are sample inefficient and computationally expensive. More advanced methods like Population Based Training that learn optimal schedules for hyperparameters instead of fixed settings can yield better results, but are also sample inefficient and computationally expensive. In this paper, we propose Hyperparameter Optimisation on the Fly (HOOF), a gradient-free algorithm that requires no more than one training run to automatically adapt the hyperparameter that affect the policy update directly through the gradient. The main idea is to use existing trajectories sampled by the policy gradient method to optimise a one-step improvement objective, yielding a sample and computationally efficient algorithm that is easy to implement. Our experimental results across multiple domains and algorithms show that using HOOF to learn these hyperparameter schedules leads to faster learning with improved performance.
****

Fast Structured Decoding for Sequence Models
Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, Zhihong Deng
Autoregressive sequence models achieve state-of-the-art performance in domains like machine translation. However, due to the autoregressive factorization nature, these models suffer from heavy latency during inference. Recently, non-autoregressive sequence models were proposed to speed up the inference time. However, these models assume that the decoding process of each token is conditionally independent of others. Such a generation process sometimes makes the output sentence inconsistent, and thus the learned non-autoregressive models could only achieve inferior accuracy compared to their autoregressive counterparts. To improve then decoding consistency and reduce the inference cost at the same time, we propose to incorporate a structured inference module into the non-autoregressive models. Specifically, we design an efficient approximation for Conditional Random Fields (CRF) for non-autoregressive sequence models, and further propose a dynamic transition technique to model positional contexts in the CRF. Experiments in machine translation show that while increasing little latency (8~14ms, our model could achieve significantly better translation performance than previous non-autoregressive models on different translation datasets. In particular, for the WMT14 En-De dataset, our model obtains a BLEU score of 26.80, which largely outperforms the previous non-autoregressive baselines and is only 0.61 lower in BLEU than purely autoregressive models.
****

Efficiently escaping saddle points on manifolds
Christopher Criscitiello, 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.
****

Comparison Against Task Driven Artificial Neural Networks Reveals Functional Properties in Mouse Visual Cortex

Jianghong Shi, Eric Shea-Brown, Michael Buice
Partially inspired by features of computation in visual cortex, deep neural networks compute hierarchical representations of their inputs. While these networks have been highly successful in machine learning, it is still unclear to what extent they can aid our understanding of cortical function. Several groups have developed metrics that provide a quantitative comparison between representations computed by networks and representations measured in cortex. At the same time, neuroscience is well into an unprecedented phase of large-scale data collection, as evidenced by projects such as the Allen Brain Observatory. Despite the magnitude of these efforts, in a given experiment only a fraction of units are recorded, limiting the information available about the cortical representation. Moreover, only a finite number of stimuli can be shown to an animal over the course of a realistic experiment. These limitations raise the question of how and whether metrics that compare representations of deep networks are meaningful on these data sets. Here, we empirically quantify the capabilities and limitations of these metrics due to limited image and neuron sample spaces. We find that the comparison procedure is robust to different choices of stimuli set and the level of sub-sampling that one might expect in a large scale brain survey with thousands of neurons. Using these results, we compare the representations measured in the Allen Brain Observatory in response to natural image presentations. We show that the visual cortical areas are relatively high order representations (in that they map to deeper layers of convolutional neural networks). Furthermore, we see evidence of a broad, more parallel organization rather than a sequential hierarchy, with the primary area VisP (V1) being lower order relative to the other areas.
****
Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)
Mariya Toneva, Leila Wehbe
Neural networks models for NLP are typically implemented without the explicit encoding of language rules and yet they are able to break one performance record after another. This has generated a lot of research interest in interpreting the representations learned by these networks. We propose here a novel interpretation approach that relies on the only processing system we have that does understand language: the human brain. We use brain imaging recordings of subjects reading complex natural text to interpret word and sequence embeddings from 4 recent NLP models - ELMo, USE, BERT and Transformer-XL. We study how their representations differ across layer depth, context length, and attention type. Our results reveal differences in the context-related representations across these models. Further, in the transformer models, we find an interaction between layer depth and context length, and between layer depth and attention type. We finally hypothesize that altering BERT to better align with brain recordings would enable it to also better understand language. Probing the altered BERT using syntactic NLP tasks reveals that the model with increased brain-alignment outperforms the original model. Cognitive neuroscientists have already begun using NLP networks to study the brain, and this work closes the loop to allow the interaction between NLP and cognitive neuroscience to be a true cross-pollination.
****
Adversarial training for free!
Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, Tom Goldstein
Adversarial training, in which a network is trained on adversarial examples, is one of the few defenses against adversarial attacks that withstands strong attacks. Unfortunately, the high cost of generating strong adversarial examples makes standard adversarial training impractical on large-scale problems like ImageNet. We present an algorithm that eliminates the overhead cost of generating adversarial examples by recycling the gradient information computed when updating model parameters. Our "free" adversarial training algorithm achieves comparable robustness to PGD adversarial training on the CIFAR-10 and CIFAR-100 datasets at negligible additional cost compared to natural training, and can be 7 to 30 times f

aster than other strong adversarial training methods. Using a single workstation with 4 P100 GPUs and 2 days of runtime, we can train a robust model for the large-scale ImageNet classification task that maintains 40% accuracy against PGD attacks.
****
Guided Similarity Separation for Image Retrieval
Chundi Liu, Guangwei Yu, Maksims Volkovs, Cheng Chang, Himanshu Rai, Junwei Ma, Satya Krishna Gorti
Despite recent progress in computer vision, image retrieval remains a challenging open problem. Numerous variations such as view angle, lighting and occlusion make it difficult to design models that are both robust and efficient. Many leading methods traverse the nearest neighbor graph to exploit higher order neighbor information and uncover the highly complex underlying manifold. In this work we propose a different approach where we leverage graph convolutional networks to directly encode neighbor information into image descriptors. We further leverage ideas from clustering and manifold learning, and introduce an unsupervised loss based on pairwise separation of image similarities. Empirically, we demonstrate that our model is able to successfully learn a new descriptor space that significantly improves retrieval accuracy, while still allowing efficient inner product inference. Experiments on five public benchmarks show highly competitive performance with up to 24\% relative improvement in mAP over leading baselines. Full code for this work is available here: https://github.com/layer6ai-labs/GSS.
****
Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks
Lixin Fan, Kam Woh Ng, Chee Seng Chan
With substantial amount of time, resources and human (team) efforts invested to explore and develop successful deep neural networks (DNN), there emerges an urgent need to protect these inventions from being illegally copied, redistributed, or abused without respecting the intellectual properties of legitimate owners. Following recent progresses along this line, we investigate a number of watermark-based DNN ownership verification methods in the face of ambiguity attacks, which aim to cast doubts on the ownership verification by forging counterfeit watermarks. It is shown that ambiguity attacks pose serious threats to existing DNN watermarking methods. As remedies to the above-mentioned loophole, this paper proposes novel passport-based DNN ownership verification schemes which are both robust to network modifications and resilient to ambiguity attacks. The gist of embedding digital passports is to design and train DNN models in a way such that, the DNN inference performance of an original task will be significantly deteriorated due to forged passports. In other words, genuine passports are not only verified by looking for the predefined signatures, but also reasserted by the unyielding DNN model inference performances. Extensive experimental results justify the effectiveness of the proposed passport-based DNN ownership verification schemes. Code and models are available at https://github.com/kamwoh/DeepIPR
****
Addressing Failure Prediction by Learning Model Confidence
Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, Patrick Pérez
Assessing reliably the confidence of a deep neural net and predicting its failures is of primary importance for the practical deployment of these models. In this paper, we propose a new target criterion for model confidence, corresponding to the True Class Probability (TCP). We show how using the TCP is more suited than relying on the classic Maximum Class Probability (MCP). We provide in addition theoretical guarantees for TCP in the context of failure prediction. Since the true class is by essence unknown at test time, we propose to learn TCP criterion on the training set, introducing a specific learning scheme adapted to this context. Extensive experiments are conducted for validating the relevance of the proposed approach. We study various network architectures, small and large scale datasets for image classification and semantic segmentation. We show that our approach consistently outperforms several strong methods, from MCP to Bayesian uncertainty, as well as recent approaches specifically designed for failure predicti

on.
****
Communication-efficient Distributed SGD with Sketching
Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Vladimir braverman, Ion Stoica, 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.
****
Multivariate Sparse Coding of Nonstationary Covariances with Gaussian Processes
Rui Li
This paper studies statistical characteristics of multivariate observations with irregular changes in their covariance structures across input space. We propose a unified nonstationary modeling framework to jointly encode the observation correlations to generate a piece-wise representation with a hyper-level Gaussian process (GP) governing the overall contour of the pieces. In particular, we couple the encoding process with automatic relevance determination (ARD) to promote sparsity to account for the inherent redundancy. The hyper GP enables us to share statistical strength among the observation variables over a collection of GPs defined within the observation pieces to characterize the variables' respective local smoothness. Experiments conducted across domains show superior performances over the state-of-the-art methods.
****
Exponential Family Estimation via Adversarial Dynamics Embedding
Bo Dai, Zhen Liu, Hanjun Dai, Niao He, Arthur Gretton, Le Song, Dale Schuurmans
We present an efficient algorithm for maximum likelihood estimation (MLE) of exponential family models, with a general parametrization of the energy function that includes neural networks. We exploit the primal-dual view of the MLE with a kinetics augmented model to obtain an estimate associated with an adversarial dual sampler. To represent this sampler, we introduce a novel neural architecture, dynamics embedding, that generalizes Hamiltonian Monte-Carlo (HMC). The proposed approach inherits the flexibility of HMC while enabling tractable entropy estimation for the augmented model. By learning both a dual sampler and the primal model simultaneously, and sharing parameters between them, we obviate the requirement to design a separate sampling procedure once the model has been trained, leading to more effective learning. We show that many existing estimators, such as contrastive divergence, pseudo/composite-likelihood, score matching, minimum Stein discrepancy estimator, non-local contrastive objectives, noise-contrastive estimation, and minimum probability flow, are special cases of the proposed approach, each expressed by a different (fixed) dual sampler. An empirical investigation shows that adapting the sampler during MLE can significantly improve on state-of-the-art estimators.
****
Group Retention when Using Machine Learning in Sequential Decision Making: the Interplay between User Dynamics and Fairness
Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, mingyan liu
Machine Learning (ML) models trained on data from multiple demographic groups can inherit representation disparity (Hashimoto et al., 2018) that may exist in the data: the model may be less favorable to groups contributing less to the training process; this in turn can degrade population retention in these groups over time, and exacerbate representation disparity in the long run. In this study, we seek to understand the interplay between ML decisions and the underlying group representation, how they evolve in a sequential framework, and how the use of fairness criteria plays a role in this process. We show that the representation disparity can easily worsen over time under a natural user dynamics (arrival and departure) model when decisions are made based on a commonly used objective and fairness criteria, resulting in some groups diminishing entirely from the sample pool in the long run. It highlights the fact that fairness criteria have to be defined while taking into consideration the impact of decisions on user dynamics.

Toward this end, we explain how a proper fairness criterion can be selected based on a general user dynamics model.
****
Shallow RNN: Accurate Time-series Classification on Resource Constrained Devices
Don Dennis, Durmus Alp Emre Acar, Vikram Mandikal, Vinu Sankar Sadasivan, Venkatesh Saligrama, Harsha Vardhan Simhadri, Prateek Jain
Recurrent Neural Networks (RNNs) capture long dependencies and context, and
2 hence are the key component of typical sequential data based tasks. However, the
sequential nature of RNNs dictates a large inference cost for long sequences even if
the hardware supports parallelization. To induce long-term dependencies, and yet
admit parallelization, we introduce novel shallow RNNs. In this architecture, the
first layer splits the input sequence and runs several independent RNNs. The second
layer consumes the output of the first layer using a second RNN thus capturing
long dependencies. We provide theoretical justification for our architecture under
weak assumptions that we verify on real-world benchmarks. Furthermore, we show
that for time-series classification, our technique leads to substantially improved
inference time over standard RNNs without compromising accuracy. For example,
we can deploy audio-keyword classification on tiny Cortex M4 devices (100MHz
processor, 256KB RAM, no DSP available) which was not possible using standard
RNN models. Similarly, using SRNN in the popular Listen-Attend-Spell (LAS)
architecture for phoneme classification [4], we can reduce the lag inphoneme
classification by 10-12x while maintaining state-of-the-art accuracy.
****
Neural Networks with Cheap Differential Operators
Ricky T. Q. Chen, David K. Duvenaud
Gradients of neural networks can be computed efficiently for any architecture, but some applications require computing differential operators with higher time complexity. We describe a family of neural network architectures that allow easy access to a family of differential operators involving \emph{dimension-wise derivatives}, and we show how to modify the backward computation graph to compute them efficiently. We demonstrate the use of these operators for solving root-finding subproblems in implicit ODE solvers, exact density evaluation for continuous normalizing flows, and evaluating the Fokker-Planck equation for training stochastic differential equation models.
****
Towards Understanding the Importance of Shortcut Connections in Residual Networks
Tianyi Liu, Minshuo Chen, Mo Zhou, Simon S. Du, Enlu Zhou, Tuo Zhao
Residual Network (ResNet) is undoubtedly a milestone in deep learning.
ResNet is equipped with shortcut connections between layers, and exhibits efficient training using simple first order algorithms. Despite of the great empirical success, the reason behind is far from being well understood. In this paper, we study a two-layer non-overlapping convolutional ResNet. Training such a network requires solving a non-convex optimization problem with a spurious local optimum. We show, however, that gradient descent combined with proper normalization, avoids being trapped by the spurious local optimum, and converges to a global optimum in polynomial time, when the weight of the first layer is initialized at 0, and that of the second layer is initialized arbitrarily in a ball. Numerical experiments are provided to support our theory.
****
A Polynomial Time Algorithm for Log-Concave Maximum Likelihood via Locally Exponential Families
Brian Axelrod, Ilias Diakonikolas, Alistair Stewart, Anastasios Sidiropoulos, Gr

egory 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Towards Automatic Concept-based Explanations
Amirata Ghorbani, James Wexler, James Y. Zou, Been Kim
Interpretability has become an important topic of research as more machine learn
ing (ML) models are deployed and widely used to make important decisions.
    Most of the current explanation methods provide explanations through feature
 importance scores, which identify features that are important for each individu
al input. However, how to systematically summarize and interpret such per sample
 feature importance scores itself is challenging. In this work, we propose princ
iples and desiderata for \emph{concept} based explanation, which goes beyond per
-sample features to identify higher level human-understandable concepts that app
ly across the entire dataset. We develop a new algorithm, ACE, to automatically
extract visual concepts. Our systematic experiments demonstrate that \alg discov
ers concepts that are human-meaningful, coherent and important for the neural ne
twork's predictions.
****
Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs
Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib
Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran N
ayebi, Daniel Bear, Daniel L. Yamins, James J. DiCarlo
Deep convolutional artificial neural networks (ANNs) are the leading class of ca
ndidate models of the mechanisms of visual processing in the primate ventral str
eam. While initially inspired by brain anatomy, over the past years, these ANNs
have evolved from a simple eight-layer architecture in AlexNet to extremely deep
 and branching architectures, demonstrating increasingly better object categoriz
ation performance, yet bringing into question how brain-like they still are. In
particular, typical deep models from the machine learning community are often ha
rd to map onto the brain's anatomy due to their vast number of layers and missin
g biologically-important connections, such as recurrence. Here we demonstrate th
at better anatomical alignment to the brain and high performance on machine lear
ning as well as neuroscience measures do not have to be in contradiction. We dev
eloped CORnet-S, a shallow ANN with four anatomically mapped areas and recurrent
 connectivity, guided by Brain-Score, a new large-scale composite of neural and
behavioral benchmarks for quantifying the functional fidelity of models of the p
rimate ventral visual stream. Despite being significantly shallower than most mo
dels, CORnet-S is the top model on Brain-Score and outperforms similarly compact
 models on ImageNet. Moreover, our extensive analyses of CORnet-S circuitry vari
ants reveal that recurrence is the main predictive factor of both Brain-Score an
d ImageNet top-1 performance. Finally, we report that the temporal evolution of
the CORnet-S "IT" neural population resembles the actual monkey IT population dy
namics. Taken together, these results establish CORnet-S, a compact, recurrent A
NN, as the current best model of the primate ventral visual stream.
****
Defending Neural Backdoors via Generative Distribution Modeling
Ximing Qiao, Yukun Yang, Hai Li
Neural backdoor attack is emerging as a severe security threat to deep learning,
 while the capability of existing defense methods is limited, especially for com
plex backdoor triggers. In the work, we explore the space formed by the pixel va
lues of all possible backdoor triggers. An original trigger used by an attacker
to build the backdoored model represents only a point in the space. It then will
 be generalized into a distribution of valid triggers, all of which can influenc
e the backdoored model. Thus, previous methods that model only one point of the
trigger distribution is not sufficient. Getting the entire trigger distribution,
 e.g., via generative modeling, is a key of effective defense. However, existing
 generative modeling techniques for image generation are not applicable to the b

ackdoor scenario as the trigger distribution is completely unknown. In this work , we propose max-entropy staircase approximator (MESA) for high-dimensional sampling-free generative modeling and use it to recover the trigger distribution. We also develop a defense technique to remove the triggers from the backdoored model. Our experiments on Cifar10/100 dataset demonstrate the effectiveness of MESA in modeling the trigger distribution and the robustness of the proposed defense method.
****
Correlation clustering with local objectives
Sanchit Kalhan, Konstantin Makarychev, Timothy Zhou
Correlation Clustering is a powerful graph partitioning model that aims to cluster items based on the notion of similarity between items. An instance of the Correlation Clustering problem consists of a graph G (not necessarily complete) whose edges are labeled by a binary classifier as
similar and dissimilar. Classically, we are tasked with producing a clustering that minimizes the number of disagreements: an edge is in disagreement if it is a similar edge and is present across clusters or if it is a dissimilar edge and is present within a cluster. Define the disagreements vector to be an n dimensional vector indexed by the vertices, where the v-th index is the number of disagreements at vertex v.
****
Logarithmic Regret for Online Control
Naman Agarwal, Elad Hazan, Karan Singh
We study optimal regret bounds for control in linear dynamical systems under adversarially changing strongly convex cost functions, given the knowledge of transition dynamics. This includes several well studied and influential frameworks such as the Kalman filter and the linear quadratic regulator. State of the art methods achieve regret which scales as T^0.5, where T is the time horizon.
****
Offline Contextual Bayesian Optimization
Ian Char, Youngseog Chung, Willie Neiswanger, Kirthevasan Kandasamy, Andrew Oakleigh Nelson, Mark Boyer, Egemen Kolemen, Jeff Schneider
In black-box optimization, an agent repeatedly chooses a configuration to test, so as to find an optimal configuration.
In many practical problems of interest, one would like to optimize several systems, or tasks'', simultaneously; however, in most of these scenarios the current task is determined by nature. In this work, we explore theoffline'' case in which one is able to bypass nature and choose the next task to evaluate (e.g. via a simulator). Because some tasks may be easier to optimize and others may be more critical, it is crucial to leverage algorithms that not only consider which configurations to try next, but also which tasks to make evaluations for. In this work, we describe a theoretically grounded Bayesian optimization method to tackle this problem. We also demonstrate that if the model of the reward structure does a poor job of capturing variation in difficulty between tasks, then algorithms that actively pick tasks for evaluation may end up doing more harm than good. Following this, we show how our approach can be used for real world applications in science and engineering, including optimizing tokamak controls for nuclear fusion.
****
Transfer Anomaly Detection by Inferring Latent Domain Representations
Atsutoshi Kumagai, Tomoharu Iwata, Yasuhiro Fujiwara
We propose a method to improve the anomaly detection performance on target domains by transferring knowledge on related domains. Although anomaly labels are valuable to learn anomaly detectors, they are difficult to obtain due to their rarity.
To alleviate this problem, existing methods use anomalous and normal instances in the related domains as well as target normal instances. These methods require training on each target domain. However, this requirement can be problematic in some situations due to the high computational cost of training. The proposed method can infer the anomaly detectors for target domains without re-training by

introducing the concept of latent domain vectors, which are latent representations of the domains and are used for inferring the anomaly detectors. The latent domain vector for each domain is inferred from the set of normal instances in the domain. The anomaly score function for each domain is modeled on the basis of autoencoders, and its domain-specific property is controlled by the latent domain vector. The anomaly score function for each domain is trained so that the scores of normal instances become low and the scores of anomalies become higher than those of the normal instances, while considering the uncertainty of the latent domain vectors. When target normal instances can be used during training, the proposed method can also use them for training in a unified framework. The effectiveness of the proposed method is demonstrated through experiments using one synthetic and four real-world datasets. Especially, the proposed method without re-training outperforms existing methods with target specific training.

****

Uncertainty on Asynchronous Time Event Prediction

Marin Biloš, Bertrand Charpentier, Stephan Günnemann

Asynchronous event sequences are the basis of many applications throughout different industries. In this work, we tackle the task of predicting the next event (given a history), and how this prediction changes with the passage of time. Since at some time points (e.g. predictions far into the future) we might not be able to predict anything with confidence, capturing uncertainty in the predictions is crucial. We present two new architectures, WGP-LN and FD-Dir, modelling the evolution of the distribution on the probability simplex with time-dependent logistic normal and Dirichlet distributions. In both cases, the combination of RNNs with either Gaussian process or function decomposition allows to express rich temporal evolution of the distribution parameters, and naturally captures uncertainty. Experiments on class prediction, time prediction and anomaly detection demonstrate the high performances of our models on various datasets compared to other approaches.

****

Breaking the Glass Ceiling for Embedding-Based Classifiers for Large Output Spaces

Chuan Guo, Ali Mousavi, Xiang Wu, Daniel N. Holtmann-Rice, Satyen Kale, Sashank Reddi, Sanjiv Kumar

In extreme classification settings, embedding-based neural network models are currently not competitive with sparse linear and tree-based methods in terms of accuracy. Most prior works attribute this poor performance to the low-dimensional bottleneck in embedding-based methods. In this paper, we demonstrate that theoretically there is no limitation to using low-dimensional embedding-based methods, and provide experimental evidence that overfitting is the root cause of the poor performance of embedding-based methods. These findings motivate us to investigate novel data augmentation and regularization techniques to mitigate overfitting. To this end, we propose GLaS, a new regularizer for embedding-based neural network approaches. It is a natural generalization from the graph Laplacian and spread-out regularizers, and empirically it addresses the drawback of each regularizer alone when applied to the extreme classification setup. With the proposed techniques, we attain or improve upon the state-of-the-art on most widely tested public extreme classification datasets with hundreds of thousands of labels.

****

Faster width-dependent algorithm for mixed packing and covering LPs

Digvijay Boob, Saurabh Sawlani, Di 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.

****
Hierarchical Decision Making by Generating and Following Natural Language Instru ctions
Hengyuan Hu, Denis Yarats, Qucheng Gong, Yuandong Tian, Mike Lewis
We explore using latent natural language instructions as an expressive and compo sitional representation of complex actions for hierarchical decision making. Rather than directly selecting micro-actions, our agent first generates a latent plan in natural language, which is then executed by a separate model. We introd uce a challenging real-time strategy game environment in which the actions of a large number of units must be coordinated across long time scales. We gather a d ataset of 76 thousand pairs of instructions and executions from human play, and train instructor and executor models. Experiments show that models using natural language as a latent variable significantly outperform models that directly imi tate human actions. The compositional structure of language proves crucial to it s effectiveness for action representation. We also release our code, models and data.
****
Structured Prediction with Projection Oracles
Mathieu Blondel
We propose in this paper a general framework for deriving loss functions for str uctured prediction. In our framework, the user chooses a convex set including t he output space and provides an oracle for projecting onto that set. Given that oracle, our framework automatically generates a corresponding convex and smooth loss function. As we show, adding a projection as output layer provably makes the loss smaller. We identify the marginal polytope, the output space's convex hull, as the best convex set on which to project. However, because the projecti on onto the marginal polytope can sometimes be expensive to compute, we allow to use any convex superset instead, with potentially cheaper-to-compute projection . Since efficient projection algorithms are available for numerous convex sets, this allows us to construct loss functions for a variety of tasks. On the theo retical side, when combined with calibrated decoding, we prove that our loss fun ctions can be used as a consistent surrogate for a (potentially non-convex) targ et loss function of interest. We demonstrate our losses on label ranking, ordin al regression and multilabel classification, confirming the improved accuracy en abled by projections.
****
Sobolev Independence Criterion
Youssef Mroueh, Tom Sercu, Mattia Rigotti, Inkit Padhi, Cicero Nogueira dos Sant os
We propose the Sobolev Independence Criterion (SIC), an interpretable dependency measure between a high dimensional random variable X and a response variable Y. SIC decomposes to the sum of feature importance scores and hence can be used fo r nonlinear feature selection. SIC can be seen as a gradient regularized Integra l Probability Metric (IPM) between the joint distribution of the two random vari ables and the product of their marginals. We use sparsity inducing gradient pena lties to promote input sparsity of the critic of the IPM. In the kernel version we show that SIC can be cast as a convex optimization problem by introducing aux iliary variables that play an important role in feature selection as they are no rmalized feature importance scores. We then present a neural version of SIC wher e the critic is parameterized as a homogeneous neural network, improving its rep resentation power as well as its interpretability. We conduct experiments valida ting SIC for feature selection in synthetic and real-world experiments. We show that SIC enables reliable and interpretable discoveries, when used in conjunctio n with the holdout randomization test and knockoffs to control the False Discove ry Rate. Code is available at http://github.com/ibm/sic.
****
Accelerating Rescaled Gradient Descent: Fast Optimization of Smooth Functions
Ashia C. Wilson, Lester Mackey, Andre Wibisono
We present a family of algorithms, called descent algorithms, for optimizing con vex and non-convex functions. We also introduce a new first-order algorithm, cal

led rescaled gradient descent (RGD), and show that RGD achieves a faster converg ence rate than gradient descent provided the function is strongly smooth - a na tural generalization of the standard smoothness assumption on the objective func tion. When the objective function is convex, we present two frameworks for “acce lerating” descent methods, one in the style of Nesterov and the other in the sty le of Monteiro and Svaiter. Rescaled gradient descent can be accelerated under t he same strong smoothness assumption using both frameworks. We provide several e xamples of strongly smooth loss functions in machine learning and numerical expe riments that verify our theoretical findings.
****
Minimax Optimal Estimation of Approximate Differential Privacy on Neighboring Da tabases
Xiyang Liu, Sewoong Oh
Differential privacy has become a widely accepted notion of privacy, leading to the introduction and deployment of numerous privatization mechanisms. However, e nsuring the privacy guarantee is an error-prone process, both in designing mecha nisms and in implementing those mechanisms. Both types of errors will be greatly reduced, if we have a data-driven approach to verify privacy guarantees, from a black-box access to a mechanism. We pose it as a property estimation problem, a nd study the fundamental trade-offs involved in the accuracy in estimated privac y guarantees and the number of samples required. We introduce a novel estimator that uses polynomial approximation of a carefully chosen degree to optimally tra de-off bias and variance. With n samples, we show that this estimator achieves p erformance of a straightforward plug-in estimator with n*log(n) samples, a pheno menon referred to as effective sample size amplification. The minimax optimality of the proposed estimator is proved by comparing it to a matching fundamental l ower bound.
****
Reconciling meta-learning and continual learning with online mixtures of tasks
Ghassen Jerfel, Erin Grant, Tom Griffiths, Katherine A. Heller
Learning-to-learn or meta-learning leverages data-driven inductive bias to incre ase the efficiency of learning on a novel task. This approach encounters difficu lty when transfer is not advantageous, for instance, when tasks are considerably dissimilar or change over time. We use the connection between gradient-based me ta-learning and hierarchical Bayes to propose a Dirichlet process mixture of hie rarchical Bayesian models over the parameters of an arbitrary parametric model s uch as a neural network. In contrast to consolidating inductive biases into a si ngle set of hyperparameters, our approach of task-dependent hyperparameter selec tion better handles latent distribution shift, as demonstrated on a set of evolv ing, image-based, few-shot learning benchmarks.
****
Neural Spline Flows
Conor Durkan, Artur Bekasov, Iain Murray, George Papamakarios
A normalizing flow models a complex probability density as an invertible transfo rmation of a simple base density. Flows based on either coupling or autoregressi ve transforms both offer exact density evaluation and sampling, but rely on the parameterization of an easily invertible elementwise transformation, whose choic e determines the flexibility of these models. Building upon recent work, we prop ose a fully-differentiable module based on monotonic rational-quadratic splines, which enhances the flexibility of both coupling and autoregressive transforms w hile retaining analytic invertibility. We demonstrate that neural spline flows i mprove density estimation, variational inference, and generative modeling of ima ges.
****
Embedding Symbolic Knowledge into Deep Networks
Yaqi Xie, Ziwei Xu, Mohan S. Kankanhalli, Kuldeep S Meel, Harold Soh
In this work, we aim to leverage prior symbolic knowledge to improve the perform ance of deep models. We propose a graph embedding network that projects proposit ional formulae (and assignments) onto a manifold via an augmented Graph Convolut ional Network (GCN). To generate semantically-faithful embeddings, we develop te

chniques to recognize node heterogeneity, and semantic regularization that incorporate structural constraints into the embedding. Experiments show that our approach improves the performance of models trained to perform entailment checking and visual relation prediction. Interestingly, we observe a connection between the tractability of the propositional theory representation and the ease of embedding. Future exploration of this connection may elucidate the relationship between knowledge compilation and vector representation learning.
****

Partitioning Structure Learning for Segmented Linear Regression Trees
Xiangyu Zheng, Song Xi Chen
This paper proposes a partitioning structure learning method for segmented linear regression trees (SLRT), which assigns linear predictors over the terminal nodes. The recursive partitioning process is driven by an adaptive split selection algorithm that maximizes, at each node, a criterion function based on a conditional Kendall's τ statistic that measures the rank dependence between the regressors and the fit- ted linear residuals. Theoretical analysis shows that the split selection algorithm permits consistent identification and estimation of the unknown segments. A suffi- ciently large tree is induced by applying the split selection algorithm recursively. Then the minimal cost-complexity tree pruning procedure is applied to attain the right-sized tree, that ensures (i) the nested structure of pruned subtrees and (ii) consistent estimation to the number of segments. Implanting the SLRT as the built-in base predictor, we obtain the ensemble predictors by random forests (RF) and the proposed weighted random forests (WRF). The practical performance of the SLRT and its ensemble versions are evaluated via numerical simulations and empirical studies. The latter shows their advantageous predictive performance over a set of state-of-the-art tree-based models on well-studied public datasets.
****

Sparse Variational Inference: Bayesian Coresets from Scratch
Trevor Campbell, Boyan Beronov
The proliferation of automated inference algorithms in Bayesian statistics has provided practitioners newfound access to fast, reproducible data analysis and powerful statistical models. Designing automated methods that are also both computationally scalable and theoretically sound, however, remains a significant challenge. Recent work on Bayesian coresets takes the approach of compressing the dataset before running a standard inference algorithm, providing both scalability and guarantees on posterior approximation error. But the automation of past coreset methods is limited because they depend on the availability of a reasonable coarse posterior approximation, which is difficult to specify in practice. In the present work we remove this requirement by formulating coreset construction as sparsity-constrained variational inference within an exponential family. This perspective leads to a novel construction via greedy optimization, and also provides a unifying information-geometric view of present and past methods. The proposed Riemannian coreset construction algorithm is fully automated, requiring no problem-specific inputs aside from the probabilistic model and dataset. In addition to being significantly easier to use than past methods, experiments demonstrate that past coreset constructions are fundamentally limited by the fixed coarse posterior approximation; in contrast, the proposed algorithm is able to continually improve the coreset, providing state-of-the-art Bayesian dataset summarization with orders-of-magnitude reduction in KL divergence to the exact posterior.
****

Policy Evaluation with Latent Confounders via Optimal Balance
Andrew Bennett, Nathan Kallus
Evaluating novel contextual bandit policies using logged data is crucial in applications where exploration is costly, such as medicine. But it usually relies on the assumption of no unobserved confounders, which is bound to fail in practice. We study the question of policy evaluation when we instead have proxies for the latent confounders and develop an importance weighting method that avoids fitting a latent outcome regression model. Surprisingly, we show that there exist no

single set of weights that give unbiased evaluation regardless of outcome model, unlike the case with no unobserved confounders where density ratios are sufficient. Instead, we propose an adversarial objective and weights that minimize it, ensuring sufficient balance in the latent confounders regardless of outcome model. We develop theory characterizing the consistency of our method and tractable algorithms for it. Empirical results validate the power of our method when confounders are latent.
****
Dancing to Music
Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, Jan Kautz
Dancing to music is an instinctive move by humans. Learning to model the music-to-dance generation process is, however, a challenging problem. It requires significant efforts to measure the correlation between music and dance as one needs to simultaneously consider multiple aspects, such as style and beat of both music and dance. Additionally, dance is inherently multimodal and various following movements of a pose at any moment are equally likely. In this paper, we propose a synthesis-by-analysis learning framework to generate dance from music. In the top-down analysis phase, we decompose a dance into a series of basic dance units, through which the model learns how to move. In the bottom-up synthesis phase, the model learns how to compose a dance by combining multiple basic dancing movements seamlessly according to input music. Experimental qualitative and quantitative results demonstrate that the proposed method can synthesize realistic, diverse, style-consistent, and beat-matching dances from music.
****
Learning Hierarchical Priors in VAEs
Alexej Klushyn, Nutan Chen, Richard Kurle, Botond Cseke, Patrick van der Smagt
We propose to learn a hierarchical prior in the context of variational autoencoders to avoid the over-regularisation resulting from a standard normal prior distribution. To incentivise an informative latent representation of the data, we formulate the learning problem as a constrained optimisation problem by extending the Taming VAEs framework to two-level hierarchical models. We introduce a graph-based interpolation method, which shows that the topology of the learned latent representation corresponds to the topology of the data manifold---and present several examples, where desired properties of latent representation such as smoothness and simple explanatory factors are learned by the prior.
****
Stochastic Runge-Kutta Accelerates Langevin Monte Carlo and Beyond
Xuechen Li, Yi Wu, Lester Mackey, Murat A. Erdogdu
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.
****
From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI
Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, Michal Irani
Reconstructing observed images from fMRI brain recordings is challenging. Unfortunately, acquiring sufficient ''labeled'' pairs of {Image, fMRI} (i.e., images with their corresponding fMRI responses) to span the huge space of natural images is prohibitive for many reasons. We present a novel approach which, in addition to the scarce labeled data (training pairs), allows to train fMRI-to-image reconstruction networks also on "unlabeled" data (i.e., images without fMRI recording, and fMRI recording without images). The proposed model utilizes both an Encoder network (image-to-fMRI) and a Decoder network (fMRI-to-image). Concatenating these two networks back-to-back (Encoder-Decoder & Decoder-Encoder) allows augmenting the training data with both types of unlabeled data. Importantly, it allows training on the unlabeled test-fMRI data. This self-supervision adapts the reconstruction network to the new input test-data, despite its deviation from the s

tatistics of the scarce training data.
****
Direct Estimation of Differential Functional Graphical Models
Boxin Zhao, Y. Samuel Wang, Mladen Kolar
We consider the problem of estimating the difference between two functional undirected graphical models with shared structures. In many applications, data are naturally regarded as high-dimensional random function vectors rather than multivariate scalars. For example, electroencephalography (EEG) data are more appropriately treated as functions of time. In these problems, not only can the number of functions measured per sample be large, but each function is itself an infinite dimensional object, making estimation of model parameters challenging. We develop a method that directly estimates the difference of graphs, avoiding separate estimation of each graph, and show it is consistent in certain high-dimensional settings. We illustrate finite sample properties of our method through simulation studies. Finally, we apply our method to EEG data to uncover differences in functional brain connectivity between alcoholics and control subjects.
****
Backpropagation-Friendly Eigendecomposition
Wei Wang, Zheng Dang, Yinlin Hu, Pascal Fua, Mathieu Salzmann
Eigendecomposition (ED) is widely used in deep networks. However, the backpropagation of its results tends to be numerically unstable, whether using ED directly or approximating it with the Power Iteration method, particularly when dealing with large matrices. While this can be mitigated by partitioning the data in small and arbitrary groups, doing so has no theoretical basis and makes its impossible to exploit the power of ED to the full. In this paper, we introduce a numerically stable and differentiable approach to leveraging eigenvectors in deep networks. It can handle large matrices without requiring to split them. We demonstrate the better robustness of our approach over standard ED and PI for ZCA whitening, an alternative to batch normalization, and for PCA denoising, which we introduce as a new normalization strategy for deep networks, aiming to further denoise the network's features.
****
Reverse KL-Divergence Training of Prior Networks: Improved Uncertainty and Adversarial Robustness
Andrey Malinin, Mark Gales
Ensemble approaches for uncertainty estimation have recently been applied to the tasks of misclassification detection, out-of-distribution input detection and adversarial attack detection. Prior Networks have been proposed as an approach to efficiently emulate an ensemble of models for classification by parameterising a Dirichlet prior distribution over output distributions. These models have been shown to outperform alternative ensemble approaches, such as Monte-Carlo Dropout, on the task of out-of-distribution input detection. However, scaling Prior Networks to complex datasets with many classes is difficult using the training criteria originally proposed. This paper makes two contributions. First, we show that the appropriate training criterion for Prior Networks is the reverse KL-divergence between Dirichlet distributions. This addresses issues in the nature of the training data target distributions, enabling prior networks to be successfully trained on classification tasks with arbitrarily many classes, as well as improving out-of-distribution detection performance. Second, taking advantage of this new training criterion, this paper investigates using Prior Networks to detect adversarial attacks and proposes a generalized form of adversarial training. It is shown that the construction of successful adaptive whitebox attacks, which affect the prediction and evade detection, against Prior Networks trained on CIFAR-10 and CIFAR-100 using the proposed approach requires a greater amount of computational effort than against networks defended using standard adversarial training or MC-dropout.
****
Adversarial Fisher Vectors for Unsupervised Representation Learning
Shuangfei Zhai, Walter Talbott, Carlos Guestrin, Joshua Susskind
We examine Generative Adversarial Networks (GANs) through the lens of deep Energ

y Based Models (EBMs), with the goal of exploiting the density model that follow
s from this formulation. In contrast to a traditional view where the discriminat
or learns a constant function when reaching convergence, here we show that it ca
n provide useful information for downstream tasks, e.g., feature extraction for
classification. To be concrete, in the EBM formulation, the discriminator learn
s an unnormalized density function (i.e., the negative energy term) that charact
erizes the data manifold. We propose to evaluate both the generator and the disc
riminator by deriving corresponding Fisher Score and Fisher Information from the
 EBM. We show that by assuming that the generated examples form an estimate of t
he learned density, both the Fisher Information and the normalized Fisher Vector
s are easy to compute. We also show that we are able to derive a distance metric
 between examples and between sets of examples. We conduct experiments showing t
hat the GAN-induced Fisher Vectors demonstrate competitive performance as unsupe
rvised feature extractors for classification and perceptual similarity tasks. Co
de is available at \url{https://github.com/apple/ml-afv}.
****
Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse
James Lucas, George Tucker, Roger B. Grosse, Mohammad Norouzi
Posterior collapse in Variational Autoencoders (VAEs) with uninformative priors
arises when the variational posterior distribution closely matches the prior for
 a subset of latent variables. This paper presents a simple and intuitive explan
ation for posterior collapse through the analysis of linear VAEs and their direc
t correspondence with Probabilistic PCA (pPCA). We explain how posterior collaps
e may occur in pPCA due to local maxima in the log marginal likelihood. Unexpect
edly, we prove that the ELBO objective for the linear VAE does not introduce add
itional spurious local maxima relative to log marginal likelihood. We show furth
er that training a linear VAE with exact variational inference recovers a unique
ly identifiable global maximum corresponding to the principal component directio
ns. Empirically, we find that our linear analysis is predictive even for high-ca
pacity, non-linear VAEs and helps explain the relationship between the observati
on noise, local maxima, and posterior collapse in deep Gaussian VAEs.
****
Kernel-Based Approaches for Sequence Modeling: Connections to Neural Methods
Kevin Liang, Guoyin Wang, Yitong Li, Ricardo Henao, Lawrence Carin
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.
****
Efficient Symmetric Norm Regression via Linear Sketching
Zhao Song, Ruosong Wang, Lin Yang, Hongyang Zhang, Peilin Zhong
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.
****
Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks
Gaël Letarte, Pascal Germain, Benjamin Guedj, Francois Laviolette
We present a comprehensive study of multilayer neural networks with binary activ
ation, relying on the PAC-Bayesian theory. Our contributions are twofold: (i) we
 develop an end-to-end framework to train a binary activated deep neural network
, (ii) we provide nonvacuous PAC-Bayesian generalization bounds for binary activ
ated deep neural networks. Our results are obtained by minimizing the expected l
oss of an architecture-dependent aggregation of binary activated deep neural net
works. Our analysis inherently overcomes the fact that binary activation functio
n is non-differentiable. The performance of our approach is assessed on a thorou
gh numerical experiment protocol on real-life datasets.
****
Approximate Feature Collisions in Neural Nets
Ke Li, Tianhao Zhang, Jitendra Malik

Work on adversarial examples has shown that neural nets are surprisingly sensitive to adversarially chosen changes of small magnitude. In this paper, we show the opposite: neural nets could be surprisingly insensitive to adversarially chosen changes of large magnitude. We observe that this phenomenon can arise from the intrinsic properties of the ReLU activation function. As a result, two very different examples could share the same feature activation and therefore the same classification decision. We refer to this phenomenon as feature collision and the corresponding examples as colliding examples. We find that colliding examples are quite abundant: we empirically demonstrate the existence of polytopes of approximately colliding examples in the neighbourhood of practically any example.
****

Characterizing Bias in Classifiers using Generative Models

Daniel McDuff, Shuang Ma, Yale Song, Ashish Kapoor

Models that are learned from real-world data are often biased because the data used to train them is biased. This can propagate systemic human biases that exist and ultimately lead to inequitable treatment of people, especially minorities. To characterize bias in learned classifiers, existing approaches rely on human oracles labeling real-world examples to identify the "blind spots" of the classifiers; these are ultimately limited due to the human labor required and the finite nature of existing image examples.
We propose a simulation-based approach for interrogating classifiers using generative adversarial models in a systematic manner. We incorporate a progressive conditional generative model for synthesizing photo-realistic facial images and Bayesian Optimization for an efficient interrogation of independent facial image classification systems. We show how this approach can be used to efficiently characterize racial and gender biases in commercial systems.
****

Coresets for Archetypal Analysis

Sebastian Mair, Ulf Brefeld

Archetypal analysis represents instances as linear mixtures of prototypes (the archetypes) that lie on the boundary of the convex hull of the data. Archetypes are thus often better interpretable than factors computed by other matrix factorization techniques. However, the interpretability comes with high computational cost due to additional convexity-preserving constraints. In this paper, we propose efficient coresets for archetypal analysis. Theoretical guarantees are derived by showing that quantization errors of k-means upper bound archetypal analysis; the computation of a provable absolute-coreset can be performed in only two passes over the data. Empirically, we show that the coresets lead to improved performance on several data sets.
****

List-decodable Linear Regression

Sushrut Karmalkar, Adam Klivans, Pravesh Kothari

We give the first polynomial-time algorithm for robust regression in the list-decodable setting where an adversary can corrupt a greater than 1/2 fraction of examples.
****

Infra-slow brain dynamics as a marker for cognitive function and decline

Shagun Ajmera, Shreya Rajagopal, Razi Rehman, Devarajan Sridharan

Functional magnetic resonance imaging (fMRI) enables measuring human brain activity, in vivo. Yet, the fMRI hemodynamic response unfolds over very slow timescales (<0.1-1 Hz), orders of magnitude slower than millisecond timescales of neural spiking. It is unclear, therefore, if slow dynamics as measured with fMRI are relevant for cognitive function. We investigated this question with a novel application of Gaussian Process Factor Analysis (GPFA) and machine learning to fMRI data. We analyzed slowly sampled (1.4 Hz) fMRI data from 1000 healthy human participants (Human Connectome Project database), and applied GPFA to reduce dimensionality and extract smooth latent dynamics. GPFA dimensions with slow (<1 Hz) characteristic timescales identified, with high accuracy (>95%), the specific task that each subject was performing inside the fMRI scanner. Moreover, functional connectivity between slow GPFA latents accurately predicted inter-individual diff

erences in behavioral scores across a range of cognitive tasks. Finally, infra-slow (<0.1 Hz) latent dynamics predicted CDR (Clinical Dementia Rating) scores of individual patients, and identified patients with mild cognitive impairment (MCI) who would progress to develop Alzheimer's dementia (AD). Slow and infra-slow brain dynamics may be relevant for understanding the neural basis of cognitive function, in health and disease.
****
Fooling Neural Network Interpretations via Adversarial Model Manipulation
Juyeon Heo, Sunghwan Joo, Taesup Moon
We ask whether the neural network interpretation methods can be fooled via adversarial model manipulation, which is defined as a model fine-tuning step that aims to radically alter the explanations without hurting the accuracy of the original models, e.g., VGG19, ResNet50, and DenseNet121. By incorporating the interpretation results directly in the penalty term of the objective function for fine-tuning, we show that the state-of-the-art saliency map based interpreters, e.g., LRP, Grad-CAM, and SimpleGrad, can be easily fooled with our model manipulation. We propose two types of fooling, Passive and Active, and demonstrate such foolings generalize well to the entire validation set as well as transfer to other interpretation methods. Our results are validated by both visually showing the fooled explanations and reporting quantitative metrics that measure the deviations from the original explanations. We claim that the stability of neural network interpretation method with respect to our adversarial model manipulation is an important criterion to check for developing robust and reliable neural network interpretation method.
****
Maximum Entropy Monte-Carlo Planning
Chenjun Xiao, Ruitong Huang, Jincheng Mei, Dale Schuurmans, Martin Müller
We develop a new algorithm for online planning in large scale sequential decision problems that improves upon the worst case efficiency of UCT. The idea is to augment Monte-Carlo Tree Search (MCTS) with maximum entropy policy optimization, evaluating each search node by softmax values back-propagated from simulation. To establish the effectiveness of this approach, we first investigate the single-step decision problem, stochastic softmax bandits, and show that softmax values can be estimated at an optimal convergence rate in terms of mean squared error. We then extend this approach to general sequential decision making by developing a general MCTS algorithm, Maximum Entropy for Tree Search (MENTS). We prove that the probability of MENTS failing to identify the best decision at the root decays exponentially, which fundamentally improves the polynomial convergence rate of UCT. Our experimental results also demonstrate that MENTS is more sample efficient than UCT in both synthetic problems and Atari 2600 games.
****
Unified Sample-Optimal Property Estimation in Near-Linear Time
Yi Hao, Alon Orlitsky
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 Keypoint Learning for Guiding Class-Conditional Video Prediction
Yunji Kim, Seonghyeon Nam, In Cho, Seon Joo Kim
We propose a deep video prediction model conditioned on a single image and an action class. To generate future frames, we first detect keypoints of a moving object and predict future motion as a sequence of keypoints. The input image is then translated following the predicted keypoints sequence to compose future frames. Detecting the keypoints is central to our algorithm, and our method is trained to detect the keypoints of arbitrary objects in an unsupervised manner. Moreover, the detected keypoints of the original videos are used as pseudo-labels to learn the motion of objects. Experimental results show that our method is successfully applied to various datasets without the cost of labeling keypoints in videos. The detected keypoints are similar to human-annotated labels, and prediction

results are more realistic compared to the previous methods.
****
Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection
Xiaoyi Gu, Leman Akoglu, Alessandro Rinaldo
Nearest-neighbor (NN) procedures are well studied and widely used in both supervised and unsupervised learning problems. In this paper we are concerned with investigating the performance of NN-based methods for anomaly detection. We first show through extensive simulations that NN methods compare favorably to some of the other state-of-the-art algorithms for anomaly detection based on a set of benchmark synthetic datasets. We further consider the performance of NN methods on real datasets, and relate it to the dimensionality of the problem. Next, we analyze the theoretical properties of NN-methods for anomaly detection by studying a more general quantity called distance-to-measure (DTM), originally developed in the literature on robust geometric and topological inference. We provide finite-sample uniform guarantees for the empirical DTM and use them to derive misclassification rates for anomalous observations under various settings. In our analysis we rely on Huber's contamination model and formulate mild geometric regularity assumptions on the underlying distribution of the data.
****
Full-Gradient Representation for Neural Network Visualization
Suraj Srinivas, François Fleuret
We introduce a new tool for interpreting neural nets, namely full-gradients, which decomposes the neural net response into input sensitivity and per-neuron sensitivity components. This is the first proposed representation which satisfies two key properties: completeness and weak dependence, which provably cannot be satisfied by any saliency map-based interpretability method. Using full-gradients, we also propose an approximate saliency map representation for convolutional nets dubbed FullGrad, obtained by aggregating the full-gradient components.
****
Learnable Tree Filter for Structure-preserving Feature Transform
Lin Song, Yanwei Li, Zeming Li, Gang Yu, Hongbin Sun, Jian Sun, Nanning Zheng
Learning discriminative global features plays a vital role in semantic segmentation. And most of the existing methods adopt stacks of local convolutions or non-local blocks to capture long-range context. However, due to the absence of spatial structure preservation, these operators ignore the object details when enlarging receptive fields. In this paper, we propose the learnable tree filter to form a generic tree filtering module that leverages the structural property of minimal spanning tree to model long-range dependencies while preserving the details. Furthermore, we propose a highly efficient linear-time algorithm to reduce resource consumption. Thus, the designed modules can be plugged into existing deep neural networks conveniently. To this end, tree filtering modules are embedded to formulate a unified framework for semantic segmentation. We conduct extensive ablation studies to elaborate on the effectiveness and efficiency of the proposed method. Specifically, it attains better performance with much less overhead compared with the classic PSP block and Non-local operation under the same backbone. Our approach is proved to achieve consistent improvements on several benchmarks without bells-and-whistles. Code and models are available at https://github.com/StevenGrove/TreeFilter-Torch.
****
The Implicit Metropolis-Hastings Algorithm
Kirill Neklyudov, Evgenii Egorov, Dmitry P. Vetrov
Recent works propose using the discriminator of a GAN to filter out unrealistic samples of the generator. We generalize these ideas by introducing the implicit Metropolis-Hastings algorithm. For any implicit probabilistic model and a target distribution represented by a set of samples, implicit Metropolis-Hastings operates by learning a discriminator to estimate the density-ratio and then generating a chain of samples. Since the approximation of density ratio introduces an error on every step of the chain, it is crucial to analyze the stationary distribution of such chain. For that purpose, we present a theoretical result stating that the discriminator loss upper bounds the total variation distance between the

target distribution and the stationary distribution. Finally, we validate the proposed algorithm both for independent and Markov proposals on CIFAR-10, CelebA, ImageNet datasets.
****
Optimal Analysis of Subset-Selection Based L_p Low-Rank Approximation
Chen Dan, Hong Wang, Hongyang Zhang, Yuchen Zhou, Pradeep K. Ravikumar
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-Efficient Distributed Blockwise Momentum SGD with Error-Feedback
Shuai Zheng, Ziyue Huang, 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.
****
Coresets for Clustering with Fairness Constraints
Lingxiao Huang, Shaofeng Jiang, Nisheeth Vishnoi
In a recent work, \cite{chierichetti2017fair} studied the following ``fair'' variants of classical clustering problems such as k-means and k-median: given a set of n data points in R^d and a binary type associated to each data point, the goal is to cluster the points while ensuring that the proportion of each type in each cluster is roughly the same as its underlying proportion. Subsequent work has focused on either extending this setting to when each data point has multiple, non-disjoint sensitive types such as race and gender \cite{bera2019fair}, or to address the problem that the clustering algorithms in the above work do not scale well. The main contribution of this paper is an approach to clustering with fairness constraints that involve {\em multiple, non-disjoint} attributes, that is {\em also scalable}. Our approach is based on novel constructions of coresets: for the k-median objective, we construct an \eps-coreset of size O(\Gamma k^2 \eps^{-d}) where \Gamma is the number of distinct collections of groups that a point may belong to, and for the k-means objective, we show how to construct an \eps-coreset of size O(\Gamma k^3\eps^{-d-1}). The former result is the first known coreset construction for the fair clustering problem with the k-median objective, and the latter result removes the dependence on the size of the full dataset as in~\cite{schmidt2018fair} and generalizes it to multiple, non-disjoint attributes. Importantly, plugging our coresets into existing algorithms for fair clustering such as \cite{backurs2019scalable} results in the fastest algorithms for several cases. Empirically, we assess our approach over the \textbf{Adult} and \textbf{Bank} dataset, and show that the coreset sizes are much smaller than the full dataset; applying coresets indeed accelerates the running time of computing the fair clustering objective while ensuring that the resulting objective difference is small.
****
You Only Propagate Once: Accelerating Adversarial Training via Maximal Principle
Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, Bin Dong
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 the Hardness of Robust Classification
Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, James Worrell
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 Temporal-Difference Learning for Policy Evaluation with Per-State Uncertainty Estimates
Carlos Riquelme, Hugo Penedones, Damien Vincent, Hartmut Maennel, Sylvain Gelly, Timothy A. Mann, Andre Barreto, Gergely Neu
We consider the core reinforcement-learning problem of on-policy value function approximation from a batch of trajectory data, and focus on various issues of Temporal Difference (TD) learning and Monte Carlo (MC) policy evaluation. The two methods are known to achieve complementary bias-variance trade-off properties, with TD tending to achieve lower variance but potentially higher bias. In this paper, we argue that the larger bias of TD can be a result of the amplification of local approximation errors. We address this by proposing an algorithm that adaptively switches between TD and MC in each state, thus mitigating the propagation of errors. Our method is based on learned confidence intervals that detect biases of TD estimates. We demonstrate in a variety of policy evaluation tasks that this simple adaptive algorithm performs competitively with the best approach in hindsight, suggesting that learned confidence intervals are a powerful technique for adapting policy evaluation to use TD or MC returns in a data-driven way.
****
Hierarchical Reinforcement Learning with Advantage-Based Auxiliary Rewards
Siyuan Li, Rui Wang, Minxue Tang, Chongjie Zhang
Hierarchical Reinforcement Learning (HRL) is a promising approach to solving long-horizon problems with sparse and delayed rewards. Many existing HRL algorithms either use pre-trained low-level skills that are unadaptable, or require domain-specific information to define low-level rewards. In this paper, we aim to adapt low-level skills to downstream tasks while maintaining the generality of reward design. We propose an HRL framework which sets auxiliary rewards for low-level skill training based on the advantage function of the high-level policy. This auxiliary reward enables efficient, simultaneous learning of the high-level policy and low-level skills without using task-specific knowledge. In addition, we also theoretically prove that optimizing low-level skills with this auxiliary reward will increase the task return for the joint policy. Experimental results show that our algorithm dramatically outperforms other state-of-the-art HRL methods in Mujoco domains. We also find both low-level and high-level policies trained by our algorithm transferable.
****
Chasing Ghosts: Instruction Following as Bayesian State Tracking
Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, Stefan Lee
A visually-grounded navigation instruction can be interpreted as a sequence of expected observations and actions an agent following the correct trajectory would encounter and perform. Based on this intuition, we formulate the problem of finding the goal location in Vision-and-Language Navigation (VLN) within the framework of Bayesian state tracking - learning observation and motion models conditioned on these expectable events. Together with a mapper that constructs a semantic spatial map on-the-fly during navigation, we formulate an end-to-end differentiable Bayes filter and train it to identify the goal by predicting the most likely trajectory through the map according to the instructions. The resulting navigation policy constitutes a new approach to instruction following that explicitly models a probability distribution over states, encoding strong geometric and algorithmic priors while enabling greater explainability. Our experiments show that our approach outperforms a strong LingUNet baseline when predicting the goal location on the map. On the full VLN task, i.e. navigating to the goal location, our approach achieves promising results with less reliance on navigation constraints.
****
Near-Optimal Reinforcement Learning in Dynamic Treatment Regimes
Junzhe Zhang, Elias Bareinboim
A dynamic treatment regime (DTR) consists of a sequence of decision rules, one per stage of intervention, that dictates how to determine the treatment assignment to patients based on evolving treatments and covariates' history. These regimes are particularly effective for managing chronic disorders and is arguably one

of the key aspects towards more personalized decision-making. In this paper, we investigate the online reinforcement learning (RL) problem for selecting optimal DTRs provided that observational data is available. We develop the first adaptive algorithm that achieves near-optimal regret in DTRs in online settings, without any access to historical data. We further derive informative bounds on the system dynamics of the underlying DTR from confounded, observational data. Finally, we combine these results and develop a novel RL algorithm that efficiently learns the optimal DTR while leveraging the abundant, yet imperfect confounded observations.
****
Rethinking the CSC Model for Natural Images
Dror Simon, Michael Elad
Sparse representation with respect to an overcomplete dictionary is often used when regularizing inverse problems in signal and image processing. In recent years, the Convolutional Sparse Coding (CSC) model, in which the dictionary consists of shift invariant filters, has gained renewed interest. While this model has been successfully used in some image processing problems, it still falls behind traditional patch-based methods on simple tasks such as denoising.
  In this work we provide new insights regarding the CSC model and its capability to represent natural images, and suggest a Bayesian connection between this model and its patch-based ancestor. Armed with these observations, we suggest a novel feed-forward network that follows an MMSE approximation process to the CSC model, using strided convolutions. The performance of this supervised architecture is shown to be on par with state of the art methods while using much fewer parameters.
****
Divide and Couple: Using Monte Carlo Variational Objectives for Posterior Approximation
Justin Domke, Daniel R. Sheldon
Recent work in variational inference (VI) has used ideas from Monte Carlo estimation to obtain tighter lower bounds on the log-likelihood to be used as objectives for VI. However, there is not a systematic understanding of how optimizing different objectives relates to approximating the posterior distribution. Developing such a connection is important if the ideas are to be applied to inference—i.e., applications that require an approximate posterior and not just an approximation of the log-likelihood. Given a VI objective defined by a Monte Carlo estimator of the likelihood, we use a "divide and couple" procedure to identify augmented proposal and target distributions so that the gap between the VI objective and the log-likelihood is equal to the divergence between these distributions. Thus, after maximizing the VI objective, the augmented variational distribution may be used to approximate the posterior distribution.
****
Numerically Accurate Hyperbolic Embeddings Using Tiling-Based Models
Tao Yu, Christopher M. De Sa
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.
****
Max-value Entropy Search for Multi-Objective Bayesian Optimization
Syrine Belakaria, Aryan Deshwal, Janardhan Rao Doppa
We consider the problem of multi-objective (MO) blackbox optimization using expensive function evaluations, where the goal is to approximate the true Pareto-set of solutions by minimizing the number of function evaluations. For example, in hardware design optimization, we need to find the designs that trade-off performance, energy, and area overhead using expensive simulations. We propose a novel approach referred to as Max-value Entropy Search for Multi-objective Optimization (MESMO) to solve this problem. MESMO employs an output-space entropy based acquisition function to efficiently select the sequence of inputs for evaluation for quickly uncovering high-quality solutions.

 We also provide theoretical analysis to characterize the efficacy of MESMO. Our
 experiments on several synthetic and real-world benchmark problems show that ME
SMO consistently outperforms state-of-the-art algorithms.
****
Algorithmic Guarantees for Inverse Imaging with Untrained Network Priors
Gauri Jagatap, Chinmay Hegde
Deep neural networks as image priors have been recently introduced for problems
such as denoising, super-resolution and inpainting with promising performance
gains over hand-crafted image priors such as sparsity. Unlike learned generative
priors they do not require any training over large datasets. However, few theore
tical guarantees exist in the scope of using untrained network priors for invers
e imaging problems. We explore new applications and theory for untrained neural
network priors. Specifically, we consider the problem of solving linear inverse
problems, such as compressive sensing, as well as non-linear problems, such as c
ompressive phase retrieval. We model images to lie in the range of an untrained
deep generative network with a fixed seed. We further present a projected gradie
nt descent scheme that can be used for both compressive sensing and phase retrie
val and provide rigorous theoretical guarantees for its convergence. We also sho
w both theoretically as well as empirically that with deep neural network priors
, one can achieve better compression rates for the same image quality as compare
d to when hand crafted priors are used.
****
Categorized Bandits
Matthieu Jedor, Vianney Perchet, Jonathan Louedec
We introduce a new stochastic multi-armed bandit setting where arms are grouped
inside ``ordered'' categories. The motivating example comes from e-commerce, whe
re a customer typically has a greater appetence for items of a specific well-ide
ntified but unknown category than any other one. We introduce three concepts of
ordering between categories, inspired by stochastic dominance between random var
iables, which are gradually weaker so that more and more bandit scenarios satisf
y at least one of them. We first prove instance-dependent lower bounds on the cu
mulative regret for each of these models, indicating how the complexity of the b
andit problems increases with the generality of the ordering concept considered
. We also provide algorithms that fully leverage the structure of the model wit
h their associated theoretical guarantees. Finally, we have conducted an analys
is on real data to highlight that those ordered categories actually exist in pra
ctice.
****
Curriculum-guided Hindsight Experience Replay
Meng Fang, Tianyi Zhou, Yali Du, Lei Han, Zhengyou Zhang
In off-policy deep reinforcement learning, it is usually hard to collect suffici
ent successful experiences with sparse rewards to learn from. Hindsight experien
ce replay (HER) enables an agent to learn from failures by treating the achieved
 state of a failed experience as a pseudo goal. However, not all the failed expe
riences are equally useful to different learning stages, so it is not efficient
to replay all of them or uniform samples of them. In this paper, we propose to 1
) adaptively select the failed experiences for replay according to the proximity
 to the true goals and the curiosity of exploration over diverse pseudo goals, a
nd 2) gradually change the proportion of the goal-proximity and the diversity-ba
sed curiosity in the selection criteria: we adopt a human-like learning strategy
 that enforces more curiosity in earlier stages and changes to larger goal-proxi
mity later. This Goal-and-Curiosity-driven Curriculum Learning'' leads toCurricu
lum-guided HER (CHER)'', which adaptively and dynamically controls the explorati
on-exploitation trade-off during the learning process via hindsight experience s
election. We show that CHER improves the state of the art in challenging robotic
s environments.
****
Random Path Selection for Continual Learning
Jathushan Rajasegaran, Munawar Hayat, Salman H. Khan, Fahad Shahbaz Khan, Ling S
hao

Incremental life-long learning is a main challenge towards the long-standing goal of Artificial General Intelligence. In real-life settings, learning tasks arrive in a sequence and machine learning models must continually learn to increment already acquired knowledge. The existing incremental learning approaches fall well below the state-of-the-art cumulative models that use all training classes at once. In this paper, we propose a random path selection algorithm, called RPS-Net, that progressively chooses optimal paths for the new tasks while encouraging parameter sharing and reuse. Our approach avoids the overhead introduced by computationally expensive evolutionary and reinforcement learning based path selection strategies while achieving considerable performance gains. As an added novelty, the proposed model integrates knowledge distillation and retrospection along with the path selection strategy to overcome catastrophic forgetting. In order to maintain an equilibrium between previous and newly acquired knowledge, we propose a simple controller to dynamically balance the model plasticity. Through extensive experiments, we demonstrate that the proposed method surpasses the state-of-the-art performance on incremental learning and by utilizing parallel computation this method can run in constant time with nearly the same efficiency as a conventional deep convolutional neural network.
****
Learning Multiple Markov Chains via Adaptive Allocation
Mohammad Sadegh Talebi, Odalric-Ambrym Maillard
We study the problem of learning the transition matrices of a set of Markov chains from a single stream of observations on each chain. We assume that the Markov chains are ergodic but otherwise unknown. The learner can sample Markov chains sequentially to observe their states. The goal of the learner is to sequentially select various chains to learn transition matrices uniformly well with respect to some loss function. We introduce a notion of loss that naturally extends the squared loss for learning distributions to the case of Markov chains, and further characterize the notion of being \emph{uniformly good} in all problem instances. We present a novel learning algorithm that efficiently balances \emph{exploration} and \emph{exploitation} intrinsic to this problem, without any prior knowledge of the chains. We provide finite-sample PAC-type guarantees on the performance of the algorithm. Further, we show that our algorithm asymptotically attains an optimal loss.
****
On Single Source Robustness in Deep Fusion Models
Taewan Kim, Joydeep Ghosh
Algorithms that fuse multiple input sources benefit from both complementary and shared information. Shared information may provide robustness against faulty or noisy inputs, which is indispensable for safety-critical applications like self-driving cars. We investigate learning fusion algorithms that are robust against noise added to a single source. We first demonstrate that robustness against single source noise is not guaranteed in a linear fusion model. Motivated by this discovery, two possible approaches are proposed to increase robustness: a carefully designed loss with corresponding training algorithms for deep fusion models, and a simple convolutional fusion layer that has a structural advantage in dealing with noise. Experimental results show that both training algorithms and our fusion layer make a deep fusion-based 3D object detector robust against noise applied to a single source, while preserving the original performance on clean data.
****
GENO -- GENeric Optimization for Classical Machine Learning
Soeren Laue, Matthias Mitterreiter, Joachim Giesen
Although optimization is the longstanding, algorithmic backbone of machine learning new models still require the time-consuming implementation of new solvers. As a result, there are thousands of implementations of optimization algorithms for machine learning problems. A natural question is, if it is always necessary to implement a new solver, or is there one algorithm that is sufficient for most models. Common belief suggests that such a one-algorithm-fits-all approach cannot work, because this algorithm cannot exploit model specific structure. At least,

a generic algorithm cannot be efficient and robust on a wide variety of problems. Here, we challenge this common belief. We have designed and implemented the optimization framework GENO (GENeric Optimization) that combines a modeling language with a generic solver. GENO takes the declaration of an optimization problem and generates a solver for the specified problem class. The framework is flexible enough to encompass most of the classical machine learning problems. We show on a wide variety of classical but also some recently suggested problems that the automatically generated solvers are (1) as efficient as well engineered, specialized solvers, (2) more efficient by a decent margin than recent state-of-the-art solvers, and (3) orders of magnitude more efficient than classical modeling language plus solver approaches.
****
Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift
Stephan Rabanser, Stephan Günnemann, Zachary Lipton
We might hope that when faced with unexpected inputs, well-designed software systems would fire off warnings. Machine learning (ML) systems, however, which depend strongly on properties of their inputs (e.g. the i.i.d. assumption), tend to fail silently. This paper explores the problem of building ML systems that fail loudly, investigating methods for detecting dataset shift, identifying exemplars that most typify the shift, and quantifying shift malignancy. We focus on several datasets and various perturbations to both covariates and label distributions with varying magnitudes and fractions of data affected. Interestingly, we show that across the dataset shifts that we explore, a two-sample-testing-based approach, using pre-trained classifiers for dimensionality reduction, performs best. Moreover, we demonstrate that domain-discriminating approaches tend to be helpful for characterizing shifts qualitatively and determining if they are harmful.
****
Shadowing Properties of Optimization Algorithms
Antonio Orvieto, Aurelien Lucchi
Ordinary differential equation (ODE) models of gradient-based optimization methods can provide insights into the dynamics of learning and inspire the design of new algorithms. Unfortunately, this thought-provoking perspective is weakened by the fact that, in the worst case, the error between the algorithm steps and its ODE approximation grows exponentially with the number of iterations. In an attempt to encourage the use of continuous-time methods in optimization, we show that, if some additional regularity on the objective is assumed, the ODE representations of Gradient Descent and Heavy-ball do not suffer from the aforementioned problem, once we allow for a small perturbation on the algorithm initial condition. In the dynamical systems literature, this phenomenon is called shadowing. Our analysis relies on the concept of hyperbolicity, as well as on tools from numerical analysis.
****
Surrogate Objectives for Batch Policy Optimization in One-step Decision Making
Minmin Chen, Ramki Gummadi, Chris Harris, Dale Schuurmans
We investigate batch policy optimization for cost-sensitive classification and contextual bandits---two related tasks that obviate exploration but require generalizing from observed rewards to action selections in unseen contexts. When rewards are fully observed, we show that the expected reward objective exhibits suboptimal plateaus and exponentially many local optima in the worst case. To overcome the poor landscape, we develop a convex surrogate that is calibrated with respect to entropy regularized expected reward. We then consider the partially observed case, where rewards are recorded for only a subset of actions. Here we generalize the surrogate to partially observed data, and uncover novel objectives for batch contextual bandit training. We find that surrogate objectives remain provably sound in this setting and empirically demonstrate state-of-the-art performance.
****
No-Press Diplomacy: Modeling Multi-Agent Gameplay
Philip Paquette, Yuchen Lu, SETON STEVEN BOCCO, Max Smith, Satya O.-G., Jonathan K. Kummerfeld, Joelle Pineau, Satinder Singh, Aaron C. Courville

Diplomacy is a seven-player non-stochastic, non-cooperative game, where agents a
cquire resources through a mix of teamwork and betrayal. Reliance on trust and c
oordination makes Diplomacy the first non-cooperative multi-agent benchmark for
complex sequential social dilemmas in a rich environment. In this work, we focus
 on training an agent that learns to play the No Press version of Diplomacy wher
e there is no dedicated communication channel between players. We present DipNet
, a neural-network-based policy model for No Press Diplomacy. The model was trai
ned on a new dataset of more than 150,000 human games. Our model is trained by s
upervised learning (SL) from expert trajectories, which is then used to initiali
ze a reinforcement learning (RL) agent trained through self-play. Both the SL an
d the RL agent demonstrate state-of-the-art No Press performance by beating popu
lar rule-based bots.
****
Bayesian Batch Active Learning as Sparse Subset Approximation
Robert Pinsler, Jonathan Gordon, Eric Nalisnick, José Miguel Hernández-Lobato
Leveraging the wealth of unlabeled data produced in recent years provides great
potential for improving supervised models. When the cost of acquiring labels is
high, probabilistic active learning methods can be used to greedily select the m
ost informative data points to be labeled. However, for many large-scale problem
s standard greedy procedures become computationally infeasible and suffer from n
egligible model change. In this paper, we introduce a novel Bayesian batch activ
e learning approach that mitigates these issues. Our approach is motivated by ap
proximating the complete data posterior of the model parameters. While naive bat
ch construction methods result in correlated queries, our algorithm produces div
erse batches that enable efficient active learning at scale. We derive interpret
able closed-form solutions akin to existing active learning procedures for linea
r models, and generalize to arbitrary models using random projections. We demons
trate the benefits of our approach on several large-scale regression and classif
ication tasks.
****
On the Ineffectiveness of Variance Reduced Optimization for Deep Learning
Aaron Defazio, Leon Bottou
The application of stochastic variance reduction to optimization has shown remar
kable recent theoretical and practical success. The applicability of these techn
iques to the hard non-convex optimization problems encountered during training o
f modern deep neural networks is an open problem. We show that naive application
 of the SVRG technique and related approaches fail, and explore why.
****
Putting An End to End-to-End: Gradient-Isolated Learning of Representations
Sindy Löwe, Peter O'Connor, Bastiaan Veeling
We propose a novel deep learning method for local self-supervised representation
 learning that does not require labels nor end-to-end backpropagation but exploi
ts the natural order in data instead. Inspired by the observation that biologica
l neural networks appear to learn without backpropagating a global error signal,
 we split a deep neural network into a stack of gradient-isolated modules. Each
module is trained to maximally preserve the information of its inputs using the
InfoNCE bound from Oord et al [2018]. Despite this greedy training, we demonstra
te that each module improves upon the output of its predecessor, and that the re
presentations created by the top module yield highly competitive results on down
stream classification tasks in the audio and visual domain. The proposal enables
 optimizing modules asynchronously, allowing large-scale distributed training of
 very deep neural networks on unlabelled datasets.
****
Modular Universal Reparameterization: Deep Multi-task Learning Across Diverse Do
mains
Elliot Meyerson, Risto Miikkulainen
As deep learning applications continue to become more diverse, an interesting qu
estion arises: Can general problem solving arise from jointly learning several s
uch diverse tasks? To approach this question, deep multi-task learning is extend
ed in this paper to the setting where there is no obvious overlap between task a

rchitectures. The idea is that any set of (architecture,task) pairs can be decom posed into a set of potentially related subproblems, whose sharing is optimized by an efficient stochastic algorithm. The approach is first validated in a class ic synthetic multi-task learning benchmark, and then applied to sharing across d isparate architectures for vision, NLP, and genomics tasks. It discovers regular ities across these domains, encodes them into sharable modules, and combines the se modules systematically to improve performance in the individual tasks. The re sults confirm that sharing learned functionality across diverse domains and arch itectures is indeed beneficial, thus establishing a key ingredient for general p roblem solving in the future.
****
Decentralized Cooperative Stochastic Bandits
David Martínez-Rubio, Varun Kanade, Patrick Rebeschini
We study a decentralized cooperative stochastic multi-armed bandit problem with K arms on a network of N agents. In our model, the reward distribution of each a rm is the same for each agent and rewards are drawn independently across agents and time steps. In each round, each agent chooses an arm to play and subsequentl y sends a message to her neighbors. The goal is to minimize the overall regret o f the entire network. We design a fully decentralized algorithm that uses an acc elerated consensus procedure to compute (delayed) estimates of the average of re wards obtained by all the agents for each arm, and then uses an upper confidence bound (UCB) algorithm that accounts for the delay and error of the estimates. W e analyze the regret of our algorithm and also provide a lower bound. The regret is bounded by the optimal centralized regret plus a natural and simple term dep ending on the spectral gap of the communication matrix. Our algorithm is simpler to analyze than those proposed in prior work and it achieves better regret boun ds, while requiring less information about the underlying network. It also perfo rms better empirically.
****
Powerset Convolutional Neural Networks
Chris Wendler, Markus Püschel, Dan Alistarh
We present a novel class of convolutional neural networks (CNNs) for set functio ns, i.e., data indexed with the powerset of a finite set. The convolutions are d erived as linear, shift-equivariant functions   for various notions of shifts on set functions. The framework is fundamentally different from graph convolutions based on the Laplacian, as it provides not one but several basic shifts, one fo r each element in the ground set. Prototypical experiments with several set func tion classification tasks on synthetic datasets and on datasets derived from rea l-world hypergraphs demonstrate the potential of our new powerset CNNs.
****
Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift
Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin , Joshua Dillon, Balaji Lakshminarayanan, Jasper Snoek
Modern machine learning methods including deep learning have achieved great succ ess in predictive accuracy for supervised learning tasks, but may still fall sho rt in giving useful estimates of their predictive uncertainty. Quantifying uncer tainty is especially critical in real-world settings, which often involve input distributions that are shifted from the training distribution due to a variety o f factors including sample bias and non-stationarity. In such settings, well ca librated uncertainty estimates convey information about when a model's output sh ould (or should not) be trusted. Many probabilistic deep learning methods, incl uding Bayesian-and non-Bayesian methods, have been proposed in the literature fo r quantifying predictive uncertainty, but to our knowledge there has not previou sly been a rigorous large-scale empirical comparison of these methods under data set shift. We present a large-scale benchmark of existing state-of-the-art metho ds on classification problems and investigate the effect of dataset shift on acc uracy and calibration. We find that traditional post-hoc calibration does indee d fall short, as do several other previous methods. However, some methods that marginalize over models give surprisingly strong results across a broad spectrum

of tasks.
****
Non-Stationary Markov Decision Processes, a Worst-Case Approach using Model-Based Reinforcement Learning
Erwan Lecarpentier, Emmanuel Rachelson
This work tackles the problem of robust zero-shot planning in non-stationary stochastic environments. We study Markov Decision Processes (MDPs) evolving over time and consider Model-Based Reinforcement Learning algorithms in this setting. We make two hypotheses: 1) the environment evolves continuously with a bounded evolution rate; 2) a current model is known at each decision epoch but not its evolution. Our contribution can be presented in four points. 1) we define a specific class of MDPs that we call Non-Stationary MDPs (NSMDPs). We introduce the notion of regular evolution by making an hypothesis of Lipschitz-Continuity on the transition and reward functions w.r.t. time; 2) we consider a planning agent using the current model of the environment but unaware of its future evolution. This leads us to consider a worst-case method where the environment is seen as an adversarial agent; 3) following this approach, we propose the Risk-Averse Tree-Search (RATS) algorithm, a zero-shot Model-Based method similar to Minimax search; 4) we illustrate the benefits brought by RATS empirically and compare its performance with reference Model-Based algorithms.
****
Optimal Decision Tree with Noisy Outcomes
Su Jia, viswanath nagarajan, Fatemeh Navidi, R Ravi
A fundamental task in active learning involves performing a sequence of tests to identify an unknown hypothesis that is drawn from a known distribution. This problem, known as optimal decision tree induction, has been widely studied for decades and the asymptotically best-possible approximation algorithm has been devised for it. We study a generalization where certain test outcomes are noisy, even in the more general case when the noise is persistent, i.e., repeating the test on the scenario gives the same noisy output, disallowing simple repetition as a way to gain confidence.
We design new approximation algorithms for both the non-adaptive setting, where the test sequence must be fixed a-priori, and the adaptive setting where the test sequence depends on the outcomes of prior tests.
Previous work in the area assumed at most a constant number of noisy outcomes per test and per scenario and provided approximation ratios that were problem dependent (such as the minimum probability of a hypothesis). Our new approximation algorithms provide guarantees that are nearly best-possible and work for the general case of a large number of noisy outcomes per test or per hypothesis where the performance degrades smoothly with this number.
Our results adapt and generalize methods used for submodular ranking and stochastic set cover.
We evaluate the performance of our algorithms on two natural applications with noise: toxic chemical identification and active learning of linear classifiers. Despite our logarithmic theoretical approximation guarantees, our methods give solutions with cost very close to the information theoretic minimum, demonstrating the effectiveness of our methods.
****
Generalization Bounds for Neural Networks via Approximate Description Length
Amit Daniely, Elad Granot
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.
****
Continual Unsupervised Representation Learning
Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, Raia Hadsell
Continual learning aims to improve the ability of modern learning systems to deal with non-stationary distributions, typically by attempting to learn a series o

f tasks sequentially. Prior art in the field has largely considered supervised or reinforcement learning tasks, and often assumes full knowledge of task labels and boundaries. In this work, we propose an approach (CURL) to tackle a more general problem that we will refer to as unsupervised continual learning. The focus is on learning representations without any knowledge about task identity, and we explore scenarios when there are abrupt changes between tasks, smooth transitions from one task to another, or even when the data is shuffled.
The proposed approach performs task inference directly within the model, is able to dynamically expand to capture new concepts over its lifetime, and incorporates additional rehearsal-based techniques to deal with catastrophic forgetting.
We demonstrate the efficacy of CURL in an unsupervised learning setting with MNIST and Omniglot, where the lack of labels ensures no information is leaked about the task.
Further, we demonstrate strong performance compared to prior art in an i.i.d setting, or when adapting the technique to supervised tasks such as incremental class learning.
****
An Inexact Augmented Lagrangian Framework for Nonconvex Optimization with Nonlinear Constraints
Mehmet Fatih Sahin, Armin eftekhari, Ahmet Alacaoglu, Fabian Latorre, Volkan Cevher
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 Robust Non-Clairvoyant Dynamic Mechanism for Contextual Auctions
Yuan Deng, Sébastien Lahaie, Vahab Mirrokni
Dynamic mechanisms offer powerful techniques to improve on both revenue and efficiency by linking sequential auctions using state information, but these techniques rely on exact distributional information of the buyers’ valuations (present and future), which limits their use in learning settings. In this paper, we consider the problem of contextual auctions where the seller gradually learns a model of the buyer's valuation as a function of the context (e.g., item features) and seeks a pricing policy that optimizes revenue. Building on the concept of a bank account mechanism---a special class of dynamic mechanisms that is known to be revenue-optimal---we develop a non-clairvoyant dynamic mechanism that is robust to both estimation errors in the buyer's value distribution and strategic behavior on the part of the buyer. We then tailor its structure to achieve a policy with provably low regret against a constant approximation of the optimal dynamic mechanism in contextual auctions. Our result substantially improves on previous results that only provide revenue guarantees against static benchmarks.
****
Multiple Futures Prediction
Charlie Tang, Russ R. Salakhutdinov
Temporal prediction is critical for making intelligent and robust decisions in complex dynamic environments. Motion prediction needs to model the inherently uncertain future which often contains multiple potential outcomes, due to multi-agent interactions and the latent goals of others. Towards these goals, we introduce a probabilistic framework that efficiently learns latent variables to jointly model the multi-step future motions of agents in a scene. Our framework is data-driven and learns semantically meaningful latent variables to represent the multimodal future, without requiring explicit labels. Using a dynamic attention-based state encoder, we learn to encode the past as well as the future interactions among agents, efficiently scaling to any number of agents. Finally, our model can be used for planning via computing a conditional probability density over the trajectories of other agents given a hypothetical rollout of the ego agent. We demonstrate our algorithms by predicting vehicle trajectories of both simulated and real data, demonstrating the state-of-the-art results on several vehicle trajectory datasets.

****
Multiview Aggregation for Learning Category-Specific Shape Reconstruction
Srinath Sridhar, Davis Rempe, Julien Valentin, Bouaziz Sofien, Leonidas J. Guibas
We investigate the problem of learning category-specific 3D shape reconstruction from a variable number of RGB views of previously unobserved object instances. Most approaches for multiview shape reconstruction operate on sparse shape representations, or assume a fixed number of views. We present a method that can estimate dense 3D shape, and aggregate shape across multiple and varying number of input views. Given a single input view of an object instance, we propose a representation that encodes the dense shape of the visible object surface as well as the surface behind line of sight occluded by the visible surface. When multiple input views are available, the shape representation is designed to be aggregated into a single 3D shape using an inexpensive union operation. We train a 2D CNN to learn to predict this representation from a variable number of views (1 or more). We further aggregate multiview information by using permutation equivariant layers that promote order-agnostic view information exchange at the feature level. Experiments show that our approach is able to produce dense 3D reconstructions of objects that improve in quality as more views are added.
****
Reinforcement Learning with Convex Constraints
Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, Robert E. Schapire
In standard reinforcement learning (RL), a learning agent seeks to optimize the overall reward. However, many key aspects of a desired behavior are more naturally expressed as constraints. For instance, the designer may want to limit the use of unsafe actions, increase the diversity of trajectories to enable exploration, or approximate expert trajectories when rewards are sparse. In this paper, we propose an algorithmic scheme that can handle a wide class of constraints in RL tasks: specifically, any constraints that require expected values of some vector measurements (such as the use of an action) to lie in a convex set. This captures previously studied constraints (such as safety and proximity to an expert), but also enables new classes of constraints (such as diversity). Our approach comes with rigorous theoretical guarantees and only relies on the ability to approximately solve standard RL tasks. As a result, it can be easily adapted to work with any model-free or model-based RL. In our experiments, we show that it matches previous algorithms that enforce safety via constraints, but can also enforce new properties that these algorithms do not incorporate, such as diversity.
****
Regularization Matters: Generalization and Optimization of Neural Nets v.s. their Induced Kernel
Colin Wei, Jason D. Lee, Qiang Liu, Tengyu Ma
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 Hawkes Processes from a handful of events
Farnood Salehi, William Trouleau, Matthias Grossglauser, Patrick Thiran
Learning the causal-interaction network of multivariate Hawkes processes is a useful task in many applications. Maximum-likelihood estimation is the most common approach to solve the problem in the presence of long observation sequences. However, when only short sequences are available, the lack of data amplifies the risk of overfitting and regularization becomes critical. Due to the challenges of hyper-parameter tuning, state-of-the-art methods only parameterize regularizers by a single shared hyper-parameter, hence limiting the power of representation of the model. To solve both issues, we develop in this work an efficient algorithm based on variational expectation-maximization. Our approach is able to optimize over an extended set of hyper-parameters. It is also able to take into account the uncertainty in the model parameters by learning a posterior distribution o

ver them. Experimental results on both synthetic and real datasets show that our approach significantly outperforms state-of-the-art methods under short observation sequences.
****

MetaInit: Initializing learning by learning to initialize
Yann N. Dauphin, Samuel Schoenholz
Deep learning models frequently trade handcrafted features for deep features learned with much less human intervention using gradient descent. While this paradigm has been enormously successful, deep networks are often difficult to train and performance can depend crucially on the initial choice of parameters. In this work, we introduce an algorithm called MetaInit as a step towards automating the search for good initializations using meta-learning. Our approach is based on a hypothesis that good initializations make gradient descent easier by starting in regions that look locally linear with minimal second order effects. We formalize this notion via a quantity that we call the gradient quotient, which can be computed with any architecture or dataset. MetaInit minimizes this quantity efficiently by using gradient descent to tune the norms of the initial weight matrices. We conduct experiments on plain and residual networks and show that the algorithm can automatically recover from a class of bad initializations. MetaInit allows us to train networks and achieve performance competitive with the state-of-the-art without batch normalization or residual connections. In particular, we find that this approach outperforms normalization for networks without skip connections on CIFAR-10 and can scale to Resnet-50 models on Imagenet.
****

Time Matters in Regularizing Deep Networks: Weight Decay and Data Augmentation Affect Early Learning Dynamics, Matter Little Near Convergence
Aditya Sharad Golatkar, Alessandro Achille, Stefano Soatto
Regularization is typically understood as improving generalization by altering the landscape of local extrema to which the model eventually converges. Deep neural networks (DNNs), however, challenge this view: We show that removing regularization after an initial transient period has little effect on generalization, even if the final loss landscape is the same as if there had been no regularization. In some cases, generalization even improves after interrupting regularization. Conversely, if regularization is applied only after the initial transient, it has no effect on the final solution, whose generalization gap is as bad as if regularization never happened. This suggests that what matters for training deep networks is not just whether or how, but when to regularize. The phenomena we observe are manifest in different datasets (CIFAR-10, CIFAR-100, SVHN, ImageNet), different architectures (ResNet-18, All-CNN), different regularization methods (weight decay, data augmentation, mixup), different learning rate schedules (exponential, piece-wise constant). They collectively suggest that there is a "critical period'' for regularizing deep networks that is decisive of the final performance. More analysis should, therefore, focus on the transient rather than asymptotic behavior of learning.
****

Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation
Ke Wang, Hang Hua, Xiaojun Wan
Unsupervised text attribute transfer automatically transforms a text to alter a specific attribute (e.g. sentiment) without using any parallel data, while simultaneously preserving its attribute-independent content. The dominant approaches are trying to model the content-independent attribute separately, e.g., learning different attributes' representations or using multiple attribute-specific decoders. However, it may lead to inflexibility from the perspective of controlling the degree of transfer or transferring over multiple aspects at the same time. To address the above problems, we propose a more flexible unsupervised text attribute transfer framework which replaces the process of modeling attribute with minimal editing of latent representations based on an attribute classifier. Specifically, we first propose a Transformer-based autoencoder to learn an entangled latent representation for a discrete text, then we transform the attribute transf

er task to an optimization problem and propose the Fast-Gradient-Iterative-Modif
ication algorithm to edit the latent representation until conforming to the targ
et attribute. Extensive experimental results demonstrate that our model achieves
 very competitive performance on three public data sets. Furthermore, we also sh
ow that our model can not only control the degree of transfer freely but also al
low to transfer over multiple aspects at the same time.
****
Accurate, reliable and fast robustness evaluation
Wieland Brendel, Jonas Rauber, Matthias Kümmerer, Ivan Ustyuzhaninov, Matthias B
ethge
Throughout the past five years, the susceptibility of neural networks to minimal
 adversarial perturbations has moved from a peculiar phenomenon to a core issue
in Deep Learning. Despite much attention, however, progress towards more robust
models is significantly impaired by the difficulty of evaluating the robustness
of neural network models. Today's methods are either fast but brittle (gradient-
based attacks), or they are fairly reliable but slow (score- and decision-based
attacks). We here develop a new set of gradient-based adversarial attacks which
(a) are more reliable in the face of gradient-masking than other gradient-based
attacks, (b) perform better and are more query efficient than current state-of-t
he-art gradient-based attacks, (c) can be flexibly adapted to a wide range of ad
versarial criteria and (d) require virtually no hyperparameter tuning. These fin
dings are carefully validated across a diverse set of six different models and h
old for L0, L1, L2 and Linf in both targeted as well as untargeted scenarios. Im
plementations will soon be available in all major toolboxes (Foolbox, CleverHans
 and ART). We hope that this class of attacks will make robustness evaluations e
asier and more reliable, thus contributing to more signal in the search for more
 robust machine learning models.
****
UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constraine
d Optimization
Ali Kavis, Kfir Y. Levy, Francis Bach, Volkan Cevher
We propose a novel adaptive, accelerated algorithm for the stochastic constraine
d convex optimization setting.Our method, which is inspired by the Mirror-Prox m
ethod, \emph{simultaneously} achieves the optimal rates for smooth/non-smooth p
roblems with either deterministic/stochastic first-order oracles. This is done w
ithout any prior knowledge of the smoothness nor the noise properties of the pro
blem. To the best of our knowledge, this is the first adaptive, unified algorith
m that achieves the optimal rates in the constrained setting. We demonstrate the
 practical performance of our framework through extensive numerical experiments.
****
From Complexity to Simplicity: Adaptive ES-Active Subspaces for Blackbox Optimiz
ation
Krzysztof M. Choromanski, Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Vikas
 Sindhwani
We present a new algorithm (ASEBO) for optimizing high-dimensional blackbox func
tions. ASEBO adapts to the geometry of the function and learns optimal sets of s
ensing directions, which are used to probe it, on-the-fly. It addresses the expl
oration-exploitation trade-off of blackbox optimization with expensive blackbox
queries by continuously learning the bias of the lower-dimensional model used to
 approximate gradients of smoothings of the function via compressed sensing and
contextual bandits methods. To obtain this model, it leverages techniques from t
he emerging theory of active subspaces in a novel ES blackbox optimization conte
xt. As a result, ASEBO learns the dynamically changing intrinsic dimensionality
of the gradient space and adapts to the hardness of different stages of the opti
mization without external supervision. Consequently, it leads to more sample-eff
icient blackbox optimization than state-of-the-art algorithms. We provide theore
tical results and test ASEBO advantages over other methods empirically by evalua
ting it on the set of reinforcement learning policy optimization tasks as well a
s functions from the recently open-sourced Nevergrad library.
****

Blocking Bandits
Soumya Basu, Rajat Sen, Sujay Sanghavi, Sanjay Shakkottai
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.
****
Value Propagation for Decentralized Networked Deep Multi-agent Reinforcement Learning
Chao Qu, Shie Mannor, Huan Xu, Yuan Qi, Le Song, Junwu Xiong
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-Person Visual Imitation Learning via Decoupled Hierarchical Controller
Pratyusha Sharma, Deepak Pathak, Abhinav Gupta
We study a generalized setup for learning from demonstration to build an agent that can manipulate novel objects in unseen scenarios by looking at only a single video of human demonstration from a third-person perspective. To accomplish this goal, our agent should not only learn to understand the intent of the demonstrated third-person video in its context but also perform the intended task in its environment configuration. Our central insight is to enforce this structure explicitly during learning by decoupling what to achieve (intended task) from how to perform it (controller). We propose a hierarchical setup where a high-level module learns to generate a series of first-person sub-goals conditioned on the third-person video demonstration, and a low-level controller predicts the actions to achieve those sub-goals. Our agent acts from raw image observations without any access to the full state information. We show results on a real robotic platform using Baxter for the manipulation tasks of pouring and placing objects in a box. Project video is available at https://pathak22.github.io/hierarchical-imitation/
****
L_DMI: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise
Yilun Xu, Peng Cao, Yuqing Kong, Yizhou Wang
Accurately annotating large scale dataset is notoriously expensive both in time and in money. Although acquiring low-quality-annotated dataset can be much cheaper, it often badly damages the performance of trained models when using such dataset without particular treatment. Various methods have been proposed for learning with noisy labels. However, most methods only handle limited kinds of noise patterns, require auxiliary information or steps (e.g., knowing or estimating the noise transition matrix), or lack theoretical justification. In this paper, we propose a novel information-theoretic loss function, LDMI, for training deep neural networks robust to label noise. The core of LDMI is a generalized version of mutual information, termed Determinant based Mutual Information (DMI), which is not only information-monotone but also relatively invariant. To the best of our knowledge, LDMI is the first loss function that is provably robust to instance-independent label noise, regardless of noise pattern, and it can be applied to any existing classification neural networks straightforwardly without any auxiliary information. In addition to theoretical justification, we also empirically show that using LDMI outperforms all other counterparts in the classification task on both image dataset and natural language dataset include Fashion-MNIST, CIFAR-10, Dogs vs. Cats, MR with a variety of synthesized noise patterns and noise amounts, as well as a real-world dataset Clothing1M.
****
Learning from Bad Data via Generation
Tianyu Guo, Chang Xu, Boxin Shi, Chao Xu, Dacheng Tao
Bad training data would challenge the learning model from understanding the underlying data-generating scheme, which then increases the difficulty in achieving

satisfactory performance on unseen test data. We suppose the real data distribution lies in a distribution set supported by the empirical distribution of bad data. A worst-case formulation can be developed over this distribution set, and then be interpreted as a generation task in an adversarial manner. The connections and differences between GANs and our framework have been thoroughly discussed. We further theoretically show the influence of this generation task on learning from bad data and reveal its connection with a data-dependent regularization. Given different distance measures (\eg, Wasserstein distance or JS divergence) of distributions, we can derive different objective functions for the problem. Experimental results on different kinds of bad training data demonstrate the necessity and effectiveness of the proposed method.
****
Connective Cognition Network for Directional Visual Commonsense Reasoning
Aming Wu, Linchao Zhu, Yahong Han, Yi 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.
****
Same-Cluster Querying for Overlapping Clusters
Wasim Huleihel, Arya Mazumdar, Muriel Medard, Soumyabrata Pal
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.
****
Discriminator optimal transport
Akinori Tanaka
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.
****
Hierarchical Optimal Transport for Document Representation
Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, Justin M. Solomon
The ability to measure similarity between documents enables intelligent summarization and analysis of large corpora. Past distances between documents suffer from either an inability to incorporate semantic similarities between words or from scalability issues. As an alternative, we introduce hierarchical optimal transport as a meta-distance between documents, where documents are modeled as distributions over topics, which themselves are modeled as distributions over words. We then solve an optimal transport problem on the smaller topic space to compute a similarity score. We give conditions on the topics under which this construction defines a distance, and we relate it to the word mover's distance.
We evaluate our technique for k-NN classification and show better interpretability and scalability with comparable performance to current methods at a fraction of the cost.
****
PerspectiveNet: A Scene-consistent Image Generator for New View Synthesis in Real Indoor Environments
David Novotny, Ben Graham, Jeremy Reizenstein
Given a set of a reference RGBD views of an indoor environment, and a new viewpoint, our goal is to predict the view from that location. Prior work on new-view generation has predominantly focused on significantly constrained scenarios, typically involving artificially rendered views of isolated CAD models. Here we tackle a much more challenging version of the problem. We devise an approach that exploits known geometric properties of the scene (per-frame camera extrinsics and depth) in order to warp reference views into the new ones. The defects in the generated views are handled by a novel RGBD inpainting network, PerspectiveNet, t

hat is fine-tuned for a given scene in order to obtain images that are geometrically consistent with all the views in the scene camera system. Experiments conducted on the ScanNet and SceneNet datasets reveal performance superior to strong baselines.
****

Strategizing against No-regret Learners
Yuan Deng, Jon Schneider, Balasubramanian Sivan
How should a player who repeatedly plays a game against a no-regret learner strategize to maximize his utility? We study this question and show that under some mild assumptions, the player can always guarantee himself a utility of at least what he would get in a Stackelberg equilibrium. When the no-regret learner has only two actions, we show that the player cannot get any higher utility than the Stackelberg equilibrium utility. But when the no-regret learner has more than two actions and plays a mean-based no-regret strategy, we show that the player can get strictly higher than the Stackelberg equilibrium utility. We construct the optimal game-play for the player against a mean-based no-regret learner who has three actions. When the no-regret learner's strategy also guarantees him a no-swap regret, we show that the player cannot get anything higher than a Stackelberg equilibrium utility.
****

Sequential Experimental Design for Transductive Linear Bandits
Tanner Fiez, Lalit Jain, Kevin G. Jamieson, Lillian Ratliff
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.
****

End to end learning and optimization on graphs
Bryan Wilder, Eric Ewing, Bistra Dilkina, Milind Tambe
Real-world applications often combine learning and optimization problems on graphs. For instance, our objective may be to cluster the graph in order to detect meaningful communities (or solve other common graph optimization problems such as facility location, maxcut, and so on). However, graphs or related attributes are often only partially observed, introducing learning problems such as link prediction which must be solved prior to optimization. Standard approaches treat learning and optimization entirely separately, while recent machine learning work aims to predict the optimal solution directly from the inputs. Here, we propose an alternative decision-focused learning approach that integrates a differentiable proxy for common graph optimization problems as a layer in learned systems. The main idea is to learn a representation that maps the original optimization problem onto a simpler proxy problem that can be efficiently differentiated through. Experimental results show that our ClusterNet system outperforms both pure end-to-end approaches (that directly predict the optimal solution) and standard approaches that entirely separate learning and optimization. Code for our system is available at https://github.com/bwilder0/clusternet.
****

Efficient Meta Learning via Minibatch Proximal Update
Pan Zhou, Xiaotong Yuan, Huan Xu, Shuicheng Yan, Jiashi Feng
We address the problem of meta-learning which learns a prior over hypothesis from a sample of meta-training tasks for fast adaptation on meta-testing tasks. A particularly simple yet successful paradigm for this research is model-agnostic meta-learning (MAML). Implementation and analysis of MAML, however, can be tricky; first-order approximation is usually adopted to avoid directly computing Hessian matrix but as a result the convergence and generalization guarantees remain largely mysterious for MAML. To remedy this deficiency, in this paper we propose a minibatch proximal update based meta-learning approach for learning to efficient hypothesis transfer. The principle is to learn a prior hypothesis shared across tasks such that the minibatch risk minimization biased regularized by this prior can quickly converge to the optimal hypothesis in each training task. The prior hypothesis training model can be efficiently optimized via SGD with provable

convergence guarantees for both convex and non-convex problems. Moreover, we theoretically justify the benefit of the learnt prior hypothesis for fast adaptation to new few-shot learning tasks via minibatch proximal update. Experimental results on several few-shot regression and classification tasks demonstrate the advantages of our method over state-of-the-arts.
****
Triad Constraints for Learning Causal Structure of Latent Variables
Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, Kun Zhang
Learning causal structure from observational data has attracted much attention, and it is notoriously challenging to find the underlying structure in the presence of confounders (hidden direct common causes of two variables). In this paper, by properly leveraging the non-Gaussianity of the data, we propose to estimate the structure over latent variables with the so-called Triad constraints: we design a form of "pseudo-residual" from three variables, and show that when causal relations are linear and noise terms are non-Gaussian, the causal direction between the latent variables for the three observed variables is identifiable by checking a certain kind of independence relationship. In other words, the Triad constraints help us to locate latent confounders and determine the causal direction between them. This goes far beyond the Tetrad constraints and reveals more information about the underlying structure from non-Gaussian data. Finally, based on the Triad constraints, we develop a two-step algorithm to learn the causal structure corresponding to measurement models. Experimental results on both synthetic and real data demonstrate the effectiveness and reliability of our method.
****
Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration
Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, Peter Flach
Class probabilities predicted by most multiclass classifiers are uncalibrated, often tending towards over-confidence. With neural networks, calibration can be improved by temperature scaling, a method to learn a single corrective multiplicative factor for inputs to the last softmax layer. On non-neural models the existing methods apply binary calibration in a pairwise or one-vs-rest fashion. We propose a natively multiclass calibration method applicable to classifiers from any model class, derived from Dirichlet distributions and generalising the beta calibration method from binary classification. It is easily implemented with neural nets since it is equivalent to log-transforming the uncalibrated probabilities, followed by one linear layer and softmax. Experiments demonstrate improved probabilistic predictions according to multiple measures (confidence-ECE, classwise-ECE, log-loss, Brier score) across a wide range of datasets and classifiers. Parameters of the learned Dirichlet calibration map
provide insights to the biases in the uncalibrated model.
****
Curvilinear Distance Metric Learning
Shuo Chen, Lei Luo, Jian Yang, Chen Gong, Jun Li, Heng Huang
Distance Metric Learning aims to learn an appropriate metric that faithfully measures the distance between two data points. Traditional metric learning methods usually calculate the pairwise distance with fixed distance functions (\emph{e.g.,}\ Euclidean distance) in the projected feature spaces. However, they fail to learn the underlying geometries of the sample space, and thus cannot exactly predict the intrinsic distances between data points. To address this issue, we first reveal that the traditional linear distance metric is equivalent to the cumulative arc length between the data pair's nearest points on the learned straight measurer lines. After that, by extending such straight lines to general curved forms, we propose a Curvilinear Distance Metric Learning (CDML) method, which adaptively learns the nonlinear geometries of the training data. By virtue of Weierstrass theorem, the proposed CDML is equivalently parameterized with a 3-order tensor, and the optimization algorithm is designed to learn the tensor parameter. Theoretical analysis is derived to guarantee the effectiveness and soundness of CDML. Extensive experiments on the synthetic and real-world datasets validate th

e superiority of our method over the state-of-the-art metric learning models.
****
Sampling Networks and Aggregate Simulation for Online POMDP Planning
Hao(Jackson) Cui, Roni Khardon
The paper introduces a new algorithm for planning in partially observable Markov decision processes (POMDP) based on the idea of aggregate simulation. The algorithm uses product distributions to approximate the belief state and shows how to build a representation graph of an approximate action-value function over belief space. The graph captures the result of simulating the model in aggregate under independence assumptions, giving a symbolic representation of the value function. The algorithm supports large observation spaces using sampling networks, a representation of the process of sampling values of observations, which is integrated into the graph representation. Following previous work in MDPs this approach enables action selection in POMDPs through gradient optimization over the graph representation. This approach complements recent algorithms for POMDPs which are based on particle representations of belief states and an explicit search for action selection. Our approach enables scaling to large factored action spaces in addition to large state spaces and observation spaces. An experimental evaluation demonstrates that the algorithm provides excellent performance relative to state of the art in large POMDP problems.
****
Robust Bi-Tempered Logistic Loss Based on Bregman Divergences
Ehsan Amid, Manfred K. K. Warmuth, Rohan Anil, Tomer Koren
We introduce a temperature into the exponential function and replace the softmax output layer of the neural networks by a high-temperature generalization. Similarly, the logarithm in the loss we use for training is replaced by a low-temperature logarithm. By tuning the two temperatures, we create loss functions that are non-convex already in the single layer case. When replacing the last layer of the neural networks by our bi-temperature generalization of the logistic loss, the training becomes more robust to noise. We visualize the effect of tuning the two temperatures in a simple setting and show the efficacy of our method on large datasets. Our methodology is based on Bregman divergences and is superior to a related two-temperature method that uses the Tsallis divergence.
****
The Parameterized Complexity of Cascading Portfolio Scheduling
Eduard Eiben, Robert Ganian, Iyad Kanj, Stefan Szeider
Cascading portfolio scheduling is a static algorithm selection strategy which uses a sample of test instances to compute an optimal ordering (a cascading schedule) of a portfolio of available algorithms. The algorithms are then applied to each future instance according to this cascading schedule, until some algorithm in the schedule succeeds. Cascading algorithm scheduling has proven to be effective in several applications, including QBF solving and the generation of ImageNet classification models.
****
Non-Asymptotic Pure Exploration by Solving Games
Rémy Degenne, Wouter M. Koolen, Pierre Ménard
Pure exploration (aka active testing) is the fundamental task of sequentially gathering information to answer a query about a stochastic environment. Good algorithms make few mistakes and take few samples. Lower bounds (for multi-armed bandit models with arms in an exponential family) reveal that the sample complexity is determined by the solution to an optimisation problem. The existing state of the art algorithms achieve asymptotic optimality by solving a plug-in estimate of that optimisation problem at each step. We interpret the optimisation problem as an unknown game, and propose sampling rules based on iterative strategies to estimate and converge to its saddle point. We apply no-regret learners to obtain the first finite confidence guarantees that are adapted to the exponential family and which apply to any pure exploration query and bandit structure. Moreover, our algorithms only use a best response oracle instead of fully solving the optimisation problem.
****

Perceiving the arrow of time in autoregressive motion
Kristof Meding, Dominik Janzing, Bernhard Schölkopf, Felix A. Wichmann
Understanding the principles of causal inference in the visual system has a long history at least since the seminal studies by Albert Michotte. Many cognitive and machine learning scientists believe that intelligent behavior requires agents to possess causal models of the world. Recent ML algorithms exploit the dependence structure of additive noise terms for inferring causal structures from observational data, e.g. to detect the direction of time series; the arrow of time. This raises the question whether the subtle asymmetries between the time directions can also be perceived by humans. Here we show that human observers can indeed discriminate forward and backward autoregressive motion with non-Gaussian additive independent noise, i.e. they appear sensitive to subtle asymmetries between the time directions. We employ a so-called frozen noise paradigm enabling us to compare human performance with four different algorithms on a trial-by-trial basis: A causal inference algorithm exploiting the dependence structure of additive noise terms, a neurally inspired network, a Bayesian ideal observer model as well as a simple heuristic. Our results suggest that all human observers use similar cues or strategies to solve the arrow of time motion discrimination task, but the human algorithm is significantly different from the three machine algorithms we compared it to. In fact, our simple heuristic appears most similar to our human observers.
****

SySCD: A System-Aware Parallel Coordinate Descent Algorithm
Nikolas Ioannou, Celestine Mendler-Dünner, Thomas Parnell
In this paper we propose a novel parallel stochastic coordinate descent (SCD) algorithm with convergence guarantees that exhibits strong scalability. We start by studying a state-of-the-art parallel implementation of SCD and identify scalability as well as system-level performance bottlenecks of the respective implementation. We then take a principled approach to develop a new SCD variant which is designed to avoid the identified system bottlenecks, such as limited scaling due to coherence traffic of model sharing across threads, and inefficient CPU cache accesses. Our proposed system-aware parallel coordinate descent algorithm (SySCD) scales to many cores and across numa nodes, and offers a consistent bottom line speedup in training time of up to x12 compared to an optimized asynchronous parallel SCD algorithm and up to x42, compared to state-of-the-art GLM solvers (scikit-learn, Vowpal Wabbit, and H2O) on a range of datasets and multi-core CPU architectures.
****

Noise-tolerant fair classification
Alex Lamy, Ziyuan Zhong, Aditya K. Menon, Nakul Verma
Fairness-aware learning involves designing algorithms that do not discriminate with respect to some sensitive feature (e.g., race or gender). Existing work on the problem operates under the assumption that the sensitive feature available in one's training sample is perfectly reliable. This assumption may be violated in many real-world cases: for example, respondents to a survey may choose to conceal or obfuscate their group identity out of fear of potential discrimination. This poses the question of whether one can still learn fair classifiers given noisy sensitive features. In this paper, we answer the question in the affirmative: we show that if one measures fairness using the mean-difference score, and sensitive features are subject to noise from the mutually contaminated learning model, then owing to a simple identity we only need to change the desired fairness-tolerance. The requisite tolerance can be estimated by leveraging existing noise-rate estimators from the label noise literature. We finally show that our procedure is empirically effective on two case-studies involving sensitive feature censoring.
****

Decentralized sketching of low rank matrices
Rakshith Sharma Srinivasa, Kiryung Lee, Marius Junge, Justin Romberg
We address a low-rank matrix recovery problem where each column of a rank-r matrix X of size (d1,d2) is compressed beyond the point of recovery to size L with L

<< d1. Leveraging the joint structure between the columns, we propose a method
to recover the matrix to within an epsilon relative error in the Frobenius norm
from a total of O(r(d1 + d2)\log^6(d1 + d2)/\epsilon^2) observations. This guara
ntee holds uniformly for all incoherent matrices of rank r. In our method, we pr
opose to use a novel matrix norm called the mixed-norm along with the maximum l2
 norm of the columns to design a novel convex relaxation for low-rank recovery t
hat is tailored to our observation model. We also show that our proposed mixed-n
orm, the standard nuclear norm, and the max-norm are particular instances of con
vex regularization of low-rankness via tensor norms. Finally, we provide a scala
ble ADMM algorithm for the mixed-norm based method and demonstrate its empirical
 performance via large-scale simulations.
****
Saccader: Improving Accuracy of Hard Attention Models for Vision
Gamaleldin Elsayed, Simon Kornblith, 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
Private Testing of Distributions via Sample Permutations
Maryam Aliakbarpour, Ilias Diakonikolas, Daniel Kane, Ronitt Rubinfeld
Statistical tests are at the heart of many scientific tasks.
To validate their hypothesis, researchers in medical and
social sciences use individuals' data. The sensitivity of
participants' data requires the design of statistical tests that
ensure the privacy of the individuals in the most efficient way.
In this paper, we use the framework of property testing to design algorithms to
test the properties of the distribution that the data is drawn from with respect
 to differential privacy.
In particular, we investigate testing two fundamental properties of distribution
s: (1) testing the equivalence of two distributions when we have unequal number
s of
samples from the two distributions.
(2) Testing independence of two random variables.
In both cases, we show that our testers achieve near optimal sample complexity (
up to logarithmic factors).
Moreover, our dependence on the privacy parameter is an additive term, which ind
icates that differential privacy can be obtained
in most regimes of parameters for free.
****
NeurVPS: Neural Vanishing Point Scanning via Conic Convolution
Yichao Zhou, Haozhi Qi, Jingwei Huang, Yi Ma
We present a simple yet effective end-to-end trainable deep network with geometr
y-inspired convolutional operators for detecting vanishing points in images. Tra
ditional convolutional neural networks rely on aggregating edge features and do
not have mechanisms to directly exploit the geometric properties of vanishing po
ints as the intersections of parallel lines. In this work, we identify a canonic
al conic space in which the neural network can effectively compute the global ge
ometric information of vanishing points locally, and we propose a novel operator
 named conic convolution that can be implemented as regular convolutions in this
 space. This new operator explicitly enforces feature extractions and aggregatio
ns along the structural lines and yet has the same number of parameters as the r
egular 2D convolution. Our extensive experiments on both synthetic and real-worl
d datasets show that the proposed operator significantly improves the performanc
e of vanishing point detection over traditional methods. The code and dataset ha
ve been made publicly available at https://github.com/zhou13/neurvps.
****
Estimating Entropy of Distributions in Constant Space
Jayadev Acharya, Sourbh Bhadane, Piotr Indyk, Ziteng Sun
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.
****
Selecting Optimal Decisions via Distributionally Robust Nearest-Neighbor Regress
ion
Ruidi Chen, Ioannis Paschalidis
This paper develops a prediction-based prescriptive model for optimal decision
making that (i) predicts the outcome under each action using a robust
nonlinear model, and (ii) adopts a randomized prescriptive policy determined
by the predicted outcomes. The predictive model combines a new regularized
regression technique, which was developed using Distributionally Robust
Optimization (DRO) with an ambiguity set constructed from the Wasserstein
metric, with the K-Nearest Neighbors (K-NN) regression, which helps to
capture the nonlinearity embedded in the data. We show theoretical results
that guarantee the out-of-sample performance of the predictive model, and
prove the optimality of the randomized policy in terms of the expected true
future outcome. We demonstrate the proposed methodology on a hypertension
dataset, showing that our prescribed treatment leads to a larger reduction in
the systolic blood pressure compared to a series of alternatives. A clinically
meaningful threshold level used to activate the randomized policy is also
derived under a sub-Gaussian assumption on the predicted outcome.
****
Exploiting Local and Global Structure for Point Cloud Semantic Segmentation with
 Contextual Point Representations
Xu Wang, Jingming He, Lin Ma
In this paper, we propose one novel model for point cloud semantic segmentation,
which exploits both the local and global structures within the point cloud based
 onthe contextual point representations. Specifically, we enrich each point repr
esen-tation by performing one novel gated fusion on the point itself and its con
textualpoints. Afterwards, based on the enriched representation, we propose one
novelgraph pointnet module, relying on the graph attention block to dynamically
com-pose and update each point representation within the local point cloud struc
ture.Finally, we resort to the spatial-wise and channel-wise attention strategie
s to exploitthe point cloud global structure and thereby yield the resulting sem
antic label foreach point. Extensive results on the public point cloud database
s, namely theS3DIS and ScanNet datasets, demonstrate the effectiveness of our pr
oposed model,outperforming the state-of-the-art approaches. Our code for this pa
per is available at https://github.com/fly519/ELGS.
****
Heterogeneous Graph Learning for Visual Commonsense Reasoning
Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, Nong Xiao
Visual commonsense reasoning task aims at leading the research field into solvin
g cognition-level reasoning with the ability to predict correct answers and mean
while providing convincing reasoning paths, resulting in three sub-tasks i.e., Q
->A, QA->R and Q->AR. It poses great challenges over the proper semantic alignme
nt between vision and linguistic domains and knowledge reasoning to generate per
suasive reasoning paths. Existing works either resort to a powerful end-to-end n
etwork that cannot produce interpretable reasoning paths or solely explore intra
-relationship of visual objects (homogeneous graph) while ignoring the cross-dom
ain semantic alignment among visual concepts and linguistic words. In this paper
, we propose a new Heterogeneous Graph Learning (HGL) framework for seamlessly i
ntegrating the intra-graph and inter-graph reasoning in order to bridge the visi
on and language domain. Our HGL consists of a primal vision-to-answer heterogene
ous graph (VAHG) module and a dual question-to-answer heterogeneous graph (QAHG)
 module to interactively refine reasoning paths for semantic agreement. Moreover
, our HGL integrates a contextual voting module to exploit a long-range visual c
ontext for better global reasoning. Experiments on the large-scale Visual Common
sense Reasoning benchmark demonstrate the superior performance of our proposed m
odules on three tasks (improving 5% accuracy on Q->A, 3.5% on QA->R, 5.8% on Q->

AR).
****
Memory Efficient Adaptive Optimization
Rohan Anil, Vineet Gupta, Tomer Koren, Yoram Singer
Adaptive gradient-based optimizers such as Adagrad and Adam are crucial for achi
eving state-of-the-art performance in machine translation and language modeling.
 However, these methods maintain second-order statistics for each parameter, thu
s introducing significant memory overheads that restrict the size of the model b
eing used as well as the number of examples in a mini-batch. We describe an effe
ctive and flexible adaptive optimization method with greatly reduced memory over
head. Our method retains the benefits of per-parameter adaptivity while allowing
 significantly larger models and batch sizes. We give convergence guarantees for
 our method, and demonstrate its effectiveness in training very large translatio
n and language models with up to 2-fold speedups compared to the state-of-the-ar
t.
****
Conformal Prediction Under Covariate Shift
Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, Aaditya Ramdas
We extend conformal prediction methodology beyond the case of exchangeable data.
 In particular, we show that a weighted version of conformal prediction can be u
sed to compute distribution-free prediction intervals for problems in which the
test and training covariate distributions differ, but the likelihood ratio betwe
en the two distributions is known---or, in practice, can be estimated accurately
 from a set of unlabeled data (test covariate points). Our weighted extension of
 conformal prediction also applies more broadly, to settings in which the data s
atisfies a certain weighted notion of exchangeability. We discuss other potenti
al applications of our new conformal methodology, including latent variable and
missing data problems.
****
Adapting Neural Networks for the Estimation of Treatment Effects
Claudia Shi, David Blei, Victor Veitch
This paper addresses the use of neural networks for the estimation of treatment
effects from observational data. Generally, estimation proceeds in two stages. F
irst, we ■t models for the expected outcome and the probability of treatment (pr
opensity score). Second, we plug these ■tted models into a downstream estimator.
 Neural networks are a natural choice for the models in the ■rst step. Our quest
ion is: how can we adapt the design and training of the neural networks used in
this ■rst step in order to improve the quality of the ■nal estimate of the treat
ment effect? We propose two adaptations based on insights from the statistical l
iterature on the estimation of treatment effects. The ■rst is a new architecture
, the Dragonnet, that exploits the suf■ciency of the propensity score for estima
tion adjustment. The second is a regularization procedure, targeted regularizati
on, that induces a bias towards models that have non-parametrically optimal asym
ptotic properties ‘out-of-the-box’. Studies on benchmark datasets for causal inf
erence show these adaptations outperform existing methods.
****
Solving graph compression via optimal transport
Vikas Garg, Tommi Jaakkola
We propose a new approach to graph compression by appeal to optimal transport. T
he transport problem is seeded with prior information about node importance, att
ributes, and edges in the graph. The transport formulation can be setup for eith
er directed or undirected graphs, and its dual characterization is cast in terms
 of distributions over the nodes. The compression pertains to the support of nod
e distributions and makes the problem challenging to solve directly. To this end
, we introduce Boolean relaxations and specify conditions under which these rela
xations are exact. The relaxations admit algorithms with provably fast convergen
ce. Moreover, we provide an exact O(d log d) algorithm for the subproblem of pro
jecting a d-dimensional vector to transformed simplex constraints. Our method ou
tperforms state-of-the-art compression methods on graph classification.
****

Optimal Sampling and Clustering in the Stochastic Block Model
Se-Young Yun, Alexandre Proutiere
This paper investigates the design of joint adaptive sampling and clustering algorithms in networks whose structure follows the celebrated Stochastic Block Model (SBM). To extract hidden clusters, the interaction between edges (pairs of nodes) may be sampled sequentially, in an adaptive manner. After gathering samples, the learner returns cluster estimates. We derive information-theoretical upper bounds on the cluster recovery rate. These bounds actually reveal the optimal sequential edge sampling strategy, and interestingly, the latter does not depend on the sampling budget, but on the parameters of the SBM only. We devise a joint sampling and clustering algorithm matching the recovery rate upper bounds. The algorithm initially uses a fraction of the sampling budget to estimate the SBM parameters, and to learn the optimal sampling strategy. This strategy then guides the remaining sampling process, which confers the optimality of the algorithm. We show both analytically and numerically that adaptive edge sampling yields important improvements over random sampling (traditionally used in the SBM analysis). For example, we prove that adaptive sampling significantly enlarges the region of the SBM parameters where asymptotically exact cluster recovery is feasible.
****
Neural Shuffle-Exchange Networks - Sequence Processing in O(n log n) Time
Karlis Freivalds, Em■ls Ozoli■š, Agris Šostaks
A key requirement in sequence to sequence processing is the modeling of long range dependencies. To this end, a vast majority of the state-of-the-art models use attention mechanism which is of O(n^2) complexity that leads to slow execution for long sequences.
****
Incremental Scene Synthesis
Benjamin Planche, Xuejian Rong, Ziyan Wu, Srikrishna Karanam, Harald Kosch, YingLi Tian, Jan Ernst, ANDREAS HUTTER
We present a method to incrementally generate complete 2D or 3D scenes with the following properties: (a) it is globally consistent at each step according to a learned scene prior, (b) real observations of a scene can be incorporated while observing global consistency, (c) unobserved regions can be hallucinated locally in consistence with previous observations, hallucinations and global priors, and (d) hallucinations are statistical in nature, i.e., different scenes can be generated from the same observations. To achieve this, we model the virtual scene, where an active agent at each step can either perceive an observed part of the scene or generate a local hallucination. The latter can be interpreted as the agent's expectation at this step through the scene and can be applied to autonomous navigation. In the limit of observing real data at each point, our method converges to solving the SLAM problem. It can otherwise sample entirely imagined scenes from prior distributions. Besides autonomous agents, applications include problems where large data is required for building robust real-world applications, but few samples are available. We demonstrate efficacy on various 2D as well as 3D data.
****
Computing Linear Restrictions of Neural Networks
Matthew Sotoudeh, Aditya V. Thakur
A linear restriction of a function is the same function with its domain restricted to points on a given line. This paper addresses the problem of computing a succinct representation for a linear restriction of a piecewise-linear neural network. This primitive, which we call ExactLine, allows us to exactly characterize the result of applying the network to all of the infinitely many points on a line. In particular, ExactLine computes a partitioning of the given input line segment such that the network is affine on each partition. We present an efficient algorithm for computing ExactLine for networks that use ReLU, MaxPool, batch normalization, fully-connected, convolutional, and other layers, along with several applications. First, we show how to exactly determine decision boundaries of an ACAS Xu neural network, providing significantly improved confidence in the results compared to prior work that sampled finitely many points in the input space.

Next, we demonstrate how to exactly compute integrated gradients, which are comm only used for neural network attributions, allowing us to show that the prior he uristic-based methods had relative errors of 25-45% and show that a better sampl ing method can achieve higher accuracy with less computation. Finally, we use Ex actLine to empirically falsify the core assumption behind a well-known hypothesi s about adversarial examples, and in the process identify interesting properties of adversarially-trained networks.
****
Markov Random Fields for Collaborative Filtering
Harald Steck
In this paper, we model the dependencies among the items that are recommended to a user in a collaborative-filtering problem via a Gaussian Markov Random Field (MRF). We build upon Besag's auto-normal parameterization and pseudo-likelihood, which not only enables computationally efficient learning, but also connects th e areas of MRFs and sparse inverse covariance estimation with autoencoders and n eighborhood models, two successful approaches in collaborative filtering. We pro pose a novel approximation for learning sparse MRFs, where the trade-off between recommendation-accuracy and training-time can be controlled. At only a small fr action of the training-time compared to various baselines, including deep nonlin ear models, the proposed approach achieved competitive ranking-accuracy on all t hree well-known data-sets used in our experiments, and notably a 20% gain in acc uracy on the data-set with the largest number of items.
****
Limiting Extrapolation in Linear Approximate Value Iteration
Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, Emma Brunskill
We study linear approximate value iteration (LAVI) with a generative model. Whil e linear models may accurately represent the optimal value function using a few parameters, several empirical and theoretical studies show the combination of le ast-squares projection with the Bellman operator may be expansive, thus leading LAVI to amplify errors over iterations and eventually diverge. We introduce an a lgorithm that approximates value functions by combining Q-values estimated at a set of \textit{anchor} states. Our algorithm tries to balance the generalization and compactness of linear methods with the small amplification of errors typica l of interpolation methods. We prove that if the features at any state can be re presented as a convex combination of features at the anchor points, then errors are propagated linearly over iterations (instead of exponentially) and our metho d achieves a polynomial sample complexity bound in the horizon and the number of anchor points. These findings are confirmed in preliminary simulations in a num ber of simple problems where a traditional least-square LAVI method diverges.
****
Regularized Weighted Low Rank Approximation
Frank Ban, David Woodruff, Richard 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Structured Graph Learning Via Laplacian Spectral Constraints
Sandeep Kumar, Jiaxi Ying, Jose Vinicius de Miranda Cardoso, Daniel Palomar
Learning a graph with a specific structure is essential for interpretability and identification of the relationships among data. But structured graph learning f rom observed samples is an NP-hard combinatorial problem. In this paper, we firs t show, for a set of important graph families it is possible to convert the comb inatorial constraints of structure into eigenvalue constraints of the graph Lapl acian matrix. Then we introduce a unified graph learning framework lying at the integration of the spectral properties of the Laplacian matrix with Gaussian gra phical modeling, which is capable of learning structures of a large class of gra ph families. The proposed algorithms are provably convergent and practically ame nable for big-data specific tasks. Extensive numerical experiments with both syn thetic and real datasets demonstrate the effectiveness of the proposed methods.

An R package containing codes for all the experimental results is submitted as a supplementary file.
****
Lookahead Optimizer: k steps forward, 1 step back
Michael Zhang, James Lucas, Jimmy Ba, Geoffrey E. Hinton
The vast majority of successful deep neural networks are trained using variants of stochastic gradient descent (SGD) algorithms. Recent attempts to improve SGD can be broadly categorized into two approaches: (1) adaptive learning rate schemes, such as AdaGrad and Adam and (2) accelerated schemes, such as heavy-ball and Nesterov momentum. In this paper, we propose a new optimization algorithm, Lookahead, that is orthogonal to these previous approaches and iteratively updates two sets of weights. Intuitively, the algorithm chooses a search direction by looking ahead at the sequence of ``fast weights" generated by another optimizer. We show that Lookahead improves the learning stability and lowers the variance of its inner optimizer with negligible computation and memory cost. We empirically demonstrate Lookahead can significantly improve the performance of SGD and Adam, even with their default hyperparameter settings on ImageNet, CIFAR-10/100, neural machine translation, and Penn Treebank.
****
Finding Friend and Foe in Multi-Agent Games
Jack Serrino, Max Kleiman-Weiner, David C. Parkes, Josh Tenenbaum
Recent breakthroughs in AI for multi-agent games like Go, Poker, and Dota, have seen great strides in recent years. Yet none of these games address the real-life challenge of cooperation in the presence of unknown and uncertain teammates. This challenge is a key game mechanism in hidden role games. Here we develop the DeepRole algorithm, a multi-agent reinforcement learning agent that we test on "The Resistance: Avalon", the most popular hidden role game. DeepRole combines counterfactual regret minimization (CFR) with deep value networks trained through self-play. Our algorithm integrates deductive reasoning into vector-form CFR to reason about joint beliefs and deduce partially observable actions. We augment deep value networks with constraints that yield interpretable representations of win probabilities. These innovations enable DeepRole to scale to the full Avalon game. Empirical game-theoretic methods show that DeepRole outperforms other hand-crafted and learned agents in five-player Avalon. DeepRole played with and against human players on the web in hybrid human-agent teams. We find that DeepRole outperforms human players as both a cooperator and a competitor.
****
Layer-Dependent Importance Sampling for Training Deep and Large Graph Convolutional Networks
Difan Zou, Ziniu Hu, Yewen Wang, Song Jiang, Yizhou Sun, Quanquan Gu
Graph convolutional networks (GCNs) have recently received wide attentions, due to their successful applications in different graph tasks and different domains. Training GCNs for a large graph, however, is still a challenge. Original full-batch GCN training requires calculating the representation of all the nodes in the graph per GCN layer, which brings in high computation and memory costs. To alleviate this issue, several sampling-based methods are proposed to train GCNs on a subset of nodes. Among them, the node-wise neighbor-sampling method recursively samples a fixed number of neighbor nodes, and thus its computation cost suffers from exponential growing neighbor size across layers; while the layer-wise importance-sampling method discards the neighbor-dependent constraints, and thus the nodes sampled across layer suffer from sparse connection problem. To deal with the above two problems, we propose a new effective sampling algorithm called LAyer-Dependent ImportancE Sampling (LADIES). Based on the sampled nodes in the upper layer, LADIES selects nodes that are in the neighborhood of these nodes and uses the constructed bipartite graph to compute the importance probability. Then, it samples a fixed number of nodes according to the probability for the whole layer, and recursively conducts such procedure per layer to construct the whole computation graph. We prove theoretically and experimentally, that our proposed sampling algorithm outperforms the previous sampling methods regarding both time and memory. Furthermore, LADIES is shown to have better generalization accuracy

than original full-batch GCN, due to its stochastic nature.
****
Self-Supervised Generalisation with Meta Auxiliary Learning
Shikun Liu, Andrew Davison, Edward Johns
Learning with auxiliary tasks can improve the ability of a primary task to gener
alise. However, this comes at the cost of manually labelling auxiliary data. We
propose a new method which automatically learns appropriate labels for an auxili
ary task, such that any supervised learning task can be improved without requiri
ng access to any further data. The approach is to train two neural networks: a l
abel-generation network to predict the auxiliary labels, and a multi-task networ
k to train the primary task alongside the auxiliary task. The loss for the label
-generation network incorporates the loss of the multi-task network, and so this
 interaction between the two networks can be seen as a form of meta learning wit
h a double gradient. We show that our proposed method, Meta AuXiliary Learning (
MAXL), outperforms single-task learning on 7 image datasets, without requiring a
ny additional data. We also show that MAXL outperforms several other baselines f
or generating auxiliary labels, and is even competitive when compared with human
-defined auxiliary labels. The self-supervised nature of our method leads to a p
romising new direction towards automated generalisation. Source code can be foun
d at \url{https://github.com/lorenmt/maxl}.
****
On Robustness of Principal Component Regression
Anish Agarwal, Devavrat Shah, Dennis Shen, Dogyoon Song
Consider the setting of Linear Regression where the observed response variables,
 in expectation, are linear functions of the p-dimensional covariates. Then to a
chieve vanishing prediction error, the number of required samples scales faster
than $p\sigma^2$, where σ^2 is a bound on the noise variance. In a high-dimensional setti
ng where p is large but the covariates admit a low-dimensional representation (s
ay r ■ p), then Principal Component Regression (PCR), cf. [36], is an effective
approach; here, the response variables are regressed with respect to the princip
al components of the covariates. The resulting number of required samples to ach
ieve vanishing prediction error now scales faster than $r\sigma^2$(■ $p\sigma^2$). Despite the t
remendous utility of PCR, its ability to handle settings with noisy, missing, an
d mixed (discrete and continuous) valued covariates is not understood and remain
s an important open challenge, cf. [24]. As the main contribution of this work,
we address this challenge by rigorously establishing that PCR is robust to noisy
, sparse, and possibly mixed valued covariates. Specifically, under PCR, vanishi
ng prediction error is achieved with the number of samples scaling as r max(σ^2,
ρ^{-4} log5(p)), where ρ denotes the fraction of observed (noisy) covariates. We es
tablish generalization error bounds on the performance of PCR, which provides a
systematic approach in selecting the correct number of components r in a data-dr
iven manner. The key to our result is a simple, but powerful equivalence between
 (i) PCR and (ii) Linear Regression with covariate pre-processing via Hard Singu
lar Value Thresholding (HSVT). From a technical standpoint, this work advances t
he state-of-the-art analysis for HSVT by establishing stronger guarantees with r
espect to the ■·■$_{2,\infty}$-error for the estimated matrix rather than the Frobenius no
rm/mean-squared error (MSE) as is commonly done in the matrix estimation / compl
etion literature.
****
Data Parameters: A New Family of Parameters for Learning a Differentiable Curric
ulum
Shreyas Saxena, Oncel Tuzel, Dennis DeCoste
Recent works have shown that learning from easier instances first can help deep
neural networks (DNNs) generalize better. However, knowing which data to present
 during different stages of training is a challenging problem. In this work, we
address
this problem by introducing data parameters. More specifically, we equip each sa
mple and class in a dataset with a learnable parameter (data parameters), which
governs their importance in the learning process. During training, at each itera
tion,

as we update the model parameters, we also update the data parameters. These upd
ates are done by gradient descent and do not require hand-crafted rules or desig
n. When applied to image classification task on CIFAR10, CIFAR100,WebVision and
ImageNet datasets, and object detection task on KITTI dataset, learning a dynami
c curriculum via data parameters leads to consistent gains, without any increase
 in model complexity or training time. When applied to a noisy dataset, the prop
osed method learns to learn from clean images and improves over the state-of-the
-art methods by 14%. To the best of our knowledge, our work is the first curricu
lum learning method to show gains on large scale image classification and detect
ion tasks.
****
One-Shot Object Detection with Co-Attention and Co-Excitation
Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, Tyng-Luh Liu
This paper aims to tackle the challenging problem of one-shot object detection.
Given a query image patch whose class label is not included in the training data
, the goal of the task is to detect all instances of the same class in a target
image. To this end, we develop a novel {\em co-attention and co-excitation} (CoA
E) framework that makes contributions in three key technical aspects. First, we
propose to use the non-local operation to explore the co-attention embodied in e
ach query-target pair and yield region proposals accounting for the one-shot sit
uation. Second, we formulate a squeeze-and-co-excitation scheme that can adaptiv
ely emphasize correlated feature channels to help uncover relevant proposals and
 eventually the target objects. Third, we design a margin-based ranking loss for
 implicitly learning a metric to predict the similarity of a region proposal to
the underlying query, no matter its class label is seen or unseen in training. T
he resulting model is therefore a two-stage detector that yields a strong baseli
ne on both VOC and MS-COCO under one-shot setting of detecting objects from both
 seen and never-seen classes.
****
Connections Between Mirror Descent, Thompson Sampling and the Information Ratio
Julian Zimmert, Tor Lattimore
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.
****
Are Anchor Points Really Indispensable in Label-Noise Learning?
Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, Masashi Sug
iyama
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.
****
SCAN: A Scalable Neural Networks Framework Towards Compact and Efficient Models
Linfeng Zhang, Zhanhong Tan, Jiebo Song, Jingwei Chen, Chenglong Bao, Kaisheng M
a
Remarkable achievements have been attained by deep neural networks in various ap
plications. However, the increasing depth and width of such models also lead to
explosive growth in both storage and computation, which has restricted the deplo
yment of deep neural networks on resource-limited edge devices. To address this
problem, we propose the so-called SCAN framework for networks training and infer
ence, which is orthogonal and complementary to existing acceleration and compres
sion methods. The proposed SCAN firstly divides neural networks into multiple se
ctions according to their depth and constructs shallow classifiers upon the inte
rmediate features of different sections. Moreover, attention modules and knowled
ge distillation are utilized to enhance the accuracy of shallow classifiers. Bas
ed on this architecture, we further propose a threshold controlled scalable infe
rence mechanism to approach human-like sample-specific inference. Experimental r
esults show that SCAN can be easily equipped on various neural networks without

any adjustment on hyper-parameters or neural networks architectures, yielding significant performance gain on CIFAR100 and ImageNet. Codes will be released on github soon.
****

Multi-Resolution Weak Supervision for Sequential Data
Paroma Varma, Frederic Sala, Shiori Sagawa, Jason Fries, Daniel Fu, Saelig Khattar, Ashwini Ramamoorthy, Ke Xiao, Kayvon Fatahalian, James Priest, Christopher Ré
Since manually labeling training data is slow and expensive, recent industrial and scientific research efforts have turned to weaker or noisier forms of supervision sources. However, existing weak supervision approaches fail to model multi-resolution sources for sequential data, like video, that can assign labels to individual elements or collections of elements in a sequence. A key challenge in weak supervision is estimating the unknown accuracies and correlations of these sources without using labeled data. Multi-resolution sources exacerbate this challenge due to complex correlations and sample complexity that scales in the length of the sequence. We propose Dugong, the first framework to model multi-resolution weak supervision sources with complex correlations to assign probabilistic labels to training data. Theoretically, we prove that Dugong, under mild conditions, can uniquely recover the unobserved accuracy and correlation parameters and use parameter sharing to improve sample complexity. Our method assigns clinician-validated labels to population-scale biomedical video repositories, helping outperform traditional supervision by 36.8 F1 points and addressing a key use case where machine learning has been severely limited by the lack of expert labeled data. On average, Dugong improves over traditional supervision by 16.0 F1 points and existing weak supervision approaches by 24.2 F1 points across several video and sensor classification tasks.
****

Smoothing Structured Decomposable Circuits
Andy Shih, Guy Van den Broeck, Paul Beame, Antoine Amarilli
We study the task of smoothing a circuit, i.e., ensuring that all children of a plus-gate mention the same variables. Circuits serve as the building blocks of state-of-the-art inference algorithms on discrete probabilistic graphical models and probabilistic programs. They are also important for discrete density estimation algorithms. Many of these tasks require the input circuit to be smooth. However, smoothing has not been studied in its own right yet, and only a trivial quadratic algorithm is known. This paper studies efficient smoothing for structured decomposable circuits. We propose a near-linear time algorithm for this task and explore lower bounds for smoothing decomposable circuits, using existing results on range-sum queries. Further, for the important case of All-Marginals, we show a more efficient linear-time algorithm. We validate experimentally the performance of our methods.
****

Bayesian Joint Estimation of Multiple Graphical Models
Lingrui Gan, Xinming Yang, Naveen Narisetty, Feng Liang
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.
****

Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion
Joan Serrà, Santiago Pascual, Carlos Segura Perales
End-to-end models for raw audio generation are a challenge, specially if they have to work with non-parallel data, which is a desirable setup in many situations. Voice conversion, in which a model has to impersonate a speaker in a recording, is one of those situations. In this paper, we propose Blow, a single-scale normalizing flow using hypernetwork conditioning to perform many-to-many voice conversion between raw audio. Blow is trained end-to-end, with non-parallel data, on a frame-by-frame basis using a single speaker identifier. We show that Blow com

pares favorably to existing flow-based architectures and other competitive baselines, obtaining equal or better performance in both objective and subjective evaluations. We further assess the impact of its main components with an ablation study, and quantify a number of properties such as the necessary amount of training data or the preference for source or target speakers.
****

Maximum Mean Discrepancy Gradient Flow

Michael Arbel, Anna Korba, Adil SALIM, Arthur Gretton

We construct a Wasserstein gradient flow of the maximum mean discrepancy (MMD) and study its convergence properties.

The MMD is an integral probability metric defined for a reproducing kernel Hilbert space (RKHS), and serves as a metric on probability measures for a sufficiently rich RKHS. We obtain conditions for convergence of the gradient flow towards a global optimum, that can be related to particle transport when optimizing neural networks.

We also propose a way to regularize this MMD flow, based on an injection of noise in the gradient. This algorithmic fix comes with theoretical and empirical evidence.

The practical implementation of the flow is straightforward, since both the MMD and its gradient have simple closed-form expressions, which can be easily estimated with samples.
****

Causal Confusion in Imitation Learning

Pim de Haan, Dinesh Jayaraman, Sergey Levine

Behavioral cloning reduces policy learning to supervised learning by training a discriminative model to predict expert actions given observations. Such discriminative models are non-causal: the training procedure is unaware of the causal structure of the interaction between the expert and the environment. We point out that ignoring causality is particularly damaging because of the distributional shift in imitation learning. In particular, it leads to a counter-intuitive "causal misidentification" phenomenon: access to more information can yield worse performance. We investigate how this problem arises, and propose a solution to combat it through targeted interventions---either environment interaction or expert queries---to determine the correct causal model. We show that causal misidentification occurs in several benchmark control domains as well as realistic driving settings, and validate our solution against DAgger and other baselines and ablations.
****

Dimensionality reduction: theoretical perspective on practical measures

Yair Bartal, Nova Fandina, Ofer Neiman

Dimensionality reduction plays a central role in real-world applications for Machine Learning, among many fields. In particular, metric dimensionality reduction where data from a general metric is mapped into low dimensional space, is often used as a first step before applying machine learning algorithms. In almost all these applications the quality of the embedding is measured by various average case criteria. Metric dimensionality reduction has also been studied in Math and TCS, within the extremely fruitful and influential field of metric embedding.

Yet, the vast majority of theoretical research has been devoted to analyzing the worst case behavior of embeddings and therefore has little relevance to practical settings. The goal of this paper is to bridge the gap between theory and practice view-points of metric dimensionality reduction, laying the foundation for a theoretical study of more practically oriented analysis. This paper can be viewed as providing a comprehensive theoretical framework addressing a line of research initiated by VL [NeuroIPS' 18] who have set the goal of analyzing different distortion measurement criteria, with the lens of Machine Learning applicability, from both theoretical and practical perspectives.

We complement their work by considering some important and vastly used average case criteria, some of which originated within the well-known Multi-Dimensional Scaling framework. While often studied in practice, no theoretical studies have thus far attempted at providing rigorous analysis of these criteria. In this pap

er we provide the first analysis of these, as well as the new distortion measure
 developed by [VL18] designed to possess Machine Learning desired properties. Mo
reover, we show that all measures considered can be adapted to possess similar q
ualities. The main consequences of our work are nearly tight bounds on the absol
ute values of all distortion criteria, as well as first approximation algorithms
 with provable guarantees.
****
MCP: Learning Composable Hierarchical Control with Multiplicative Compositional
Policies
Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, Sergey Levine
Humans are able to perform a myriad of sophisticated tasks by drawing upon skill
s acquired through prior experience. For autonomous agents to have this capabili
ty, they must be able to extract reusable skills from past experience that can b
e recombined in new ways for subsequent tasks. Furthermore, when controlling com
plex high-dimensional morphologies, such as humanoid bodies, tasks often require
 coordination of multiple skills simultaneously. Learning discrete primitives fo
r every combination of skills quickly becomes prohibitive. Composable primitives
 that can be recombined to create a large variety of behaviors can be more suita
ble for modeling this combinatorial explosion. In this work, we propose multipli
cative compositional policies (MCP), a method for learning reusable motor skills
 that can be composed to produce a range of complex behaviors. Our method factor
izes an agent's skills into a collection of primitives, where multiple primitive
s can be activated simultaneously via multiplicative composition. This flexibili
ty allows the primitives to be transferred and recombined to elicit new behavior
s as necessary for novel tasks. We demonstrate that MCP is able to extract compo
sable skills for highly complex simulated characters from pre-training tasks, su
ch as motion imitation, and then reuse these skills to solve challenging continu
ous control tasks, such as dribbling a soccer ball to a goal, and picking up an
object and transporting it to a target location.
****
Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networ
ks
Aaron Voelker, Ivana Kaji■, Chris Eliasmith
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.
****
BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Lear
ning
Andreas Kirsch, Joost van Amersfoort, Yarin Gal
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.
****
Generalized Matrix Means for Semi-Supervised Learning with Multilayer Graphs
Pedro Mercado, Francesco Tudisco, Matthias Hein
We study the task of semi-supervised learning on multilayer graphs by taking int
o account both labeled and unlabeled observations together with the information
encoded by each individual graph layer. We propose a regularizer based on the ge
neralized matrix mean, which is a one-parameter family of matrix means that incl
udes the arithmetic, geometric and harmonic means as particular cases. We analyz
e it in expectation under a Multilayer Stochastic Block Model and verify numeric
ally that it outperforms state of the art methods. Moreover, we introduce a matr
ix-free numerical scheme based on contour integral quadratures and Krylov subspa
ce solvers that scales to large sparse multilayer graphs.
****
Efficiently Estimating Erdos-Renyi Graphs with Node Differential Privacy
Jonathan Ullman, Adam Sealfon

We give a simple, computationally efficient, and node-differentially-private alg orithm for estimating the parameter of an Erdos-Renyi graph---that is, estimatin g p in a G(n,p)---with near-optimal accuracy. Our algorithm nearly matches the information-theoretically optimal exponential-time algorithm for the same proble m due to Borgs et al. (FOCS 2018). More generally, we give an optimal, computat ionally efficient, private algorithm for estimating the edge-density of any grap h whose degree distribution is concentrated in a small interval.
****

Screening Sinkhorn Algorithm for Regularized Optimal Transport
Mokhtar Z. Alaya, Maxime Berar, Gilles Gasso, Alain Rakotomamonjy
We introduce in this paper a novel strategy for efficiently approximating the Si nkhorn distance between two discrete measures. After identifying neglectable com ponents of the dual solution of the regularized Sinkhorn problem, we propose to screen those components by directly setting them at that value before entering t he Sinkhorn problem. This allows us to solve a smaller Sinkhorn problem while en suring approximation with provable guarantees. More formally, the approach is ba sed on a new formulation of dual of Sinkhorn divergence problem and on the KKT o ptimality conditions of this problem, which enable identification of dual compon ents to be screened. This new analysis leads to the Screenkhorn algorithm. We il lustrate the efficiency of Screenkhorn on complex tasks such as dimensionality r eduction and domain adaptation involving regularized optimal transport.
****

Adaptive Density Estimation for Generative Models
Thomas Lucas, Konstantin Shmelkov, Karteek Alahari, Cordelia Schmid, Jakob Verbe ek
Unsupervised learning of generative models has seen tremendous progress over rec ent years, in particular due to generative adversarial networks (GANs), variatio nal autoencoders, and flow-based models. GANs have dramatically improved sample quality, but suffer from two drawbacks: (i) they mode-drop, \ie, do not cover the full support of the train data, and (ii) they do not allow for likelihood ev aluations on held-out data. In contrast likelihood-based training encourages mo dels to cover the full support of the train data, but yields poorer samples. Th ese mutual shortcomings can in principle be addressed by training generative lat ent variable models in a hybrid adversarial-likelihood manner. However, we show that commonly made parametric assumptions create a conflict between them, makin g successful hybrid models non trivial. As a solution, we propose the use of de ep invertible transformations in the latent variable decoder. This approach all ows for likelihood computations in image space, is more efficient than fully inv ertible models, and can take full advantage of adversarial training. We show th at our model significantly improves over existing hybrid models: offering GAN-li ke samples, IS and FID scores that are competitive with fully adversarial models and improved likelihood scores.
****

Learning Deep Bilinear Transformation for Fine-grained Image Representation
Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, Jiebo Luo
Bilinear feature transformation has shown the state-of-the-art performance in le arning fine-grained image representations. However, the computational cost to le arn pairwise interactions between deep feature channels is prohibitively expensi ve, which restricts this powerful transformation to be used in deep neural netwo rks. In this paper, we propose a deep bilinear transformation (DBT) block, which can be deeply stacked in convolutional neural networks to learn fine-grained im age representations. The DBT block can uniformly divide input channels into seve ral semantic groups. As bilinear transformation can be represented by calculatin g pairwise interactions within each group, the computational cost can be heavily relieved. The output of each block is further obtained by aggregating intra-gro up bilinear features, with residuals from the entire input features. We found th at the proposed network achieves new state-of-the-art in several fine-grained im age recognition benchmarks, including CUB-Bird, Stanford-Car, and FGVC-Aircraft.
****

Learning Compositional Neural Programs with Recursive Tree Search and Planning

Thomas PIERROT, Guillaume Ligner, Scott E. Reed, Olivier Sigaud, Nicolas Perrin, Alexandre Laterre, David Kas, Karim Beguir, Nando de Freitas
We propose a novel reinforcement learning algorithm, AlphaNPI, that incorporates the strengths of Neural Programmer-Interpreters (NPI) and AlphaZero. NPI contributes structural biases in the form of modularity, hierarchy and recursion,
which are helpful to reduce sample complexity, improve generalization and increase interpretability. AlphaZero contributes powerful neural network guided search algorithms, which we augment with recursion. AlphaNPI only assumes a hierarchical program specification with sparse rewards: 1 when the program execution satisfies the specification, and 0 otherwise. This specification enables
us to overcome the need for strong supervision in the form of execution traces and consequently train NPI models effectively with reinforcement learning. The experiments show that AlphaNPI can sort as well as previous strongly supervised NPI variants. The AlphaNPI agent is also trained on a Tower of Hanoi puzzle with two disks and is shown to generalize to puzzles with an arbitrary number of disks.
The experiments also show that when deploying our neural network policies, it is advantageous to do planning with guided Monte Carlo tree search.
****
Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks
Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, George Pappas
Tight estimation of the Lipschitz constant for deep neural networks (DNNs) is useful in many applications ranging from robustness certification of classifiers to stability analysis of closed-loop systems with reinforcement learning controllers. Existing methods in the literature for estimating the Lipschitz constant suffer from either lack of accuracy or poor scalability. In this paper, we present a convex optimization framework to compute guaranteed upper bounds on the Lipschitz constant of DNNs both accurately and efficiently. Our main idea is to interpret activation functions as gradients of convex potential functions. Hence, they satisfy certain properties that can be described by quadratic constraints. This particular description allows us to pose the Lipschitz constant estimation problem as a semidefinite program (SDP). The resulting SDP can be adapted to increase either the estimation accuracy (by capturing the interaction between activation functions of different layers) or scalability (by decomposition and parallel implementation). We illustrate the utility of our approach with a variety of experiments on randomly generated networks and on classifiers trained on the MNIST and Iris datasets. In particular, we experimentally demonstrate that our Lipschitz bounds are the most accurate compared to those in the literature. We also study the impact of adversarial training methods on the Lipschitz bounds of the resulting classifiers and show that our bounds can be used to efficiently provide robustness guarantees.
****
Mo' States Mo' Problems: Emergency Stop Mechanisms from Observation
Samuel Ainsworth, Matt Barnes, Siddhartha Srinivasa
In many environments, only a relatively small subset of the complete state space is necessary in order to accomplish a given task. We develop a simple technique using emergency stops (e-stops) to exploit this phenomenon. Using e-stops significantly improves sample complexity by reducing the amount of required exploration, while retaining a performance bound that efficiently trades off the rate of convergence with a small asymptotic sub-optimality gap. We analyze the regret behavior of e-stops and present empirical results in discrete and continuous settings demonstrating that our reset mechanism can provide order-of-magnitude speedups on top of existing reinforcement learning methods.
****
Kernelized Bayesian Softmax for Text Generation
Ning Miao, Hao Zhou, Chengqi Zhao, Wenxian Shi, Lei Li
Neural models for text generation require a softmax layer with proper token embe

ddings during the decoding phase.
Most existing approaches adopt single point embedding for each token.
However, a word may have multiple senses according to different context, some of which might be distinct.
In this paper, we propose KerBS, a novel approach for learning better embeddings for text generation.
KerBS embodies two advantages:
(a) it employs a Bayesian composition of embeddings for words with multiple senses;
(b) it is adaptive to semantic variances of words and robust to rare sentence context by imposing learned kernels to capture the closeness of words (senses) in the embedding space.
Empirical studies show that KerBS significantly boosts the performance of several text generation tasks.
****
Bipartite expander Hopfield networks as self-decoding high-capacity error correcting codes
Rishidev Chaudhuri, Ila Fiete
Neural network models of memory and error correction famously include the Hopfield network, which can directly store---and error-correct through its dynamics---arbitrary N-bit patterns, but only for ~N such patterns. On the other end of the spectrum, Shannon's coding theory established that it is possible to represent exponentially many states (~e^N) using N symbols in such a way that an optimal decoder could correct all noise upto a threshold. We prove that it is possible to construct an associative content-addressable network that combines the properties of strong error correcting codes and Hopfield networks: it simultaneously possesses exponentially many stable states, these states are robust enough, with large enough basins of attraction that they can be correctly recovered despite errors in a finite fraction of all nodes, and the errors are intrinsically corrected by the network’s own dynamics. The network is a two-layer Boltzmann machine with simple neural dynamics, low dynamic-range (binary) pairwise synaptic connections, and sparse expander graph connectivity. Thus, quasi-random sparse structures---characteristic of important error-correcting codes---may provide for high-performance computation in artificial neural networks and the brain.
****
Distributional Reward Decomposition for Reinforcement Learning
Zichuan Lin, Li Zhao, Derek Yang, Tao Qin, Tie-Yan Liu, Guangwen Yang
Many reinforcement learning (RL) tasks have specific properties that can be leveraged to modify existing RL algorithms to adapt to those tasks and further improve performance, and a general class of such properties is the multiple reward channel. In those environments the full reward can be decomposed into sub-rewards obtained from different channels. Existing work on reward decomposition either requires prior knowledge of the environment to decompose the full reward, or decomposes reward without prior knowledge but with degraded performance. In this paper, we propose Distributional Reward Decomposition for Reinforcement Learning (DRDRL), a novel reward decomposition algorithm which captures the multiple reward channel structure under distributional setting. Empirically, our method captures the multi-channel structure and discovers meaningful reward decomposition, without any requirements on prior knowledge. Consequently, our agent achieves better performance than existing methods on environments with multiple reward channels.
****
Provably Global Convergence of Actor-Critic: A Case for Linear Quadratic Regulator with Ergodic Cost
Zhuoran Yang, Yongxin Chen, Mingyi Hong, Zhaoran Wang
Despite the empirical success of the actor-critic algorithm, its theoretical understanding lags behind. In a broader context, actor-critic can be viewed as an online alternating update algorithm for bilevel optimization, whose convergence is known to be fragile. To understand the instability of actor-critic, we focus on its application to linear quadratic regulators, a simple yet fundamental setti

ng of reinforcement learning. We establish a nonasymptotic convergence analysis of actor- critic in this setting. In particular, we prove that actor-critic finds a globally optimal pair of actor (policy) and critic (action-value function) at a linear rate of convergence. Our analysis may serve as a preliminary step towards a complete theoretical understanding of bilevel optimization with nonconvex subproblems, which is NP-hard in the worst case and is often solved using heuristics.
****
Fast-rate PAC-Bayes Generalization Bounds via Shifted Rademacher Processes
Jun Yang, Shengyang Sun, Daniel M. Roy
The developments of Rademacher complexity and PAC-Bayesian theory have been largely independent. One exception is the PAC-Bayes theorem of Kakade, Sridharan, and Tewari (2008), which is established via Rademacher complexity theory by viewing Gibbs classifiers as linear operators. The goal of this paper is to extend this bridge between Rademacher complexity and state-of-the-art PAC-Bayesian theory. We first demonstrate that one can match the fast rate of Catoni's PAC-Bayes bounds (Catoni, 2007) using shifted Rademacher processes (Wegkamp, 2003; Lecué and Mitchell, 2012; Zhivotovskiy and Hanneke, 2018). We then derive a new fast-rate PAC-Bayes bound in terms of the "flatness" of the empirical risk surface on which the posterior concentrates. Our analysis establishes a new framework for deriving fast-rate PAC-Bayes bounds and yields new insights on PAC-Bayesian theory.
****
DINGO: Distributed Newton-Type Method for Gradient-Norm Optimization
Rixon Crane, Fred Roosta
For optimization of a large sum of functions in a distributed computing environment, we present a novel communication efficient Newton-type algorithm that enjoys a variety of advantages over similar existing methods. Our algorithm, DINGO, is derived by optimization of the gradient's norm as a surrogate function. DINGO does not impose any specific form on the underlying functions and its application range extends far beyond convexity and smoothness. The underlying sub-problems of DINGO are simple linear least-squares, for which a plethora of efficient algorithms exist. DINGO involves a few hyper-parameters that are easy to tune and we theoretically show that a strict reduction in the surrogate objective is guaranteed, regardless of the selected hyper-parameters.
****
Deep ReLU Networks Have Surprisingly Few Activation Patterns
Boris Hanin, David Rolnick
The success of deep networks has been attributed in part to their expressivity: per parameter, deep networks can approximate a richer class of functions than shallow networks. In ReLU networks, the number of activation patterns is one measure of expressivity; and the maximum number of patterns grows exponentially with the depth. However, recent work has showed that the practical expressivity of deep networks - the functions they can learn rather than express - is often far from the theoretical maximum. In this paper, we show that the average number of activation patterns for ReLU networks at initialization is bounded by the total number of neurons raised to the input dimension. We show empirically that this bound, which is independent of the depth, is tight both at initialization and during training, even on memorization tasks that should maximize the number of activation patterns. Our work suggests that realizing the full expressivity of deep networks may not be possible in practice, at least with current methods.
****
Private Hypothesis Selection
Mark Bun, Gautam Kamath, Thomas Steinke, Steven Z. 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.
****
ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, Boris Katz

We collect a large real-world test set, ObjectNet, for object recognition with controls where object backgrounds, rotations, and imaging viewpoints are random. Most scientific experiments have controls, confounds which are removed from the data, to ensure that subjects cannot perform a task by exploiting trivial correlations in the data. Historically, large machine learning and computer vision datasets have lacked such controls. This has resulted in models that must be fine-tuned for new datasets and perform better on datasets than in real-world applications. When tested on ObjectNet, object detectors show a 40-45% drop in performance, with respect to their performance on other benchmarks, due to the controls for biases. Controls make ObjectNet robust to fine-tuning showing only small performance increases. We develop a highly automated platform that enables gathering datasets with controls by crowdsourcing image capturing and annotation. ObjectNet is the same size as the ImageNet test set (50,000 images), and by design does not come paired with a training set in order to encourage generalization. The dataset is both easier than ImageNet (objects are largely centered and unoccluded) and harder (due to the controls). Although we focus on object recognition here, data with controls can be gathered at scale using automated tools throughout machine learning to generate datasets that exercise models in new ways thus providing valuable feedback to researchers. This work opens up new avenues for research in generalizable, robust, and more human-like computer vision and in creating datasets where results are predictive of real-world performance.

****

Object landmark discovery through unsupervised adaptation

Enrique Sanchez, Georgios Tzimiropoulos

This paper proposes a method to ease the unsupervised learning of object landmark detectors. Similarly to previous methods, our approach is fully unsupervised in a sense that it does not require or make any use of annotated landmarks for the target object category. Contrary to previous works, we do however assume that a landmark detector, which has already learned a structured representation for a given object category in a fully supervised manner, is available. Under this setting, our main idea boils down to adapting the given pre-trained network to the target object categories in a fully unsupervised manner. To this end, our method uses the pre-trained network as a core which remains frozen and does not get updated during training, and learns, in an unsupervised manner, only a projection matrix to perform the adaptation to the target categories. By building upon an existing structured representation learned in a supervised manner, the optimization problem solved by our method is much more constrained with significantly less parameters to learn which seems to be important for the case of unsupervised learning. We show that our method surpasses fully unsupervised techniques trained from scratch as well as a strong baseline based on fine-tuning, and produces state-of-the-art results on several datasets. Code can be found at tiny.cc/GitHub-Unsupervised

****

Block Coordinate Regularization by Denoising

Yu Sun, Jiaming Liu, Ulugbek Kamilov

We consider the problem of estimating a vector from its noisy measurements using a prior specified only through a denoising function. Recent work on plug-and-play priors (PnP) and regularization-by-denoising (RED) has shown the state-of-the-art performance of estimators under such priors in a range of imaging tasks. In this work, we develop a new block coordinate RED algorithm that decomposes a large-scale estimation problem into a sequence of updates over a small subset of the unknown variables. We theoretically analyze the convergence of the algorithm and discuss its relationship to the traditional proximal optimization. Our analysis complements and extends recent theoretical results for RED-based estimation methods. We numerically validate our method using several denoiser priors, including those based on convolutional neural network (CNN) denoisers.

****

Neural Temporal-Difference Learning Converges to Global Optima

Qi Cai, Zhuoran Yang, Jason D. Lee, Zhaoran Wang
Temporal-difference learning (TD), coupled with neural networks, is among the most fundamental building blocks of deep reinforcement learning. However, due to the nonlinearity in value function approximation, such a coupling leads to nonconvexity and even divergence in optimization. As a result, the global convergence of neural TD remains unclear. In this paper, we prove for the first time that neural TD converges at a sublinear rate to the global optimum of the mean-squared projected Bellman error for policy evaluation. In particular, we show how such global convergence is enabled by the overparametrization of neural networks, which also plays a vital role in the empirical success of neural TD. Beyond policy evaluation, we establish the global convergence of neural (soft) Q-learning, which is further connected to that of policy gradient algorithms.
****
Learning Nearest Neighbor Graphs from Noisy Distance Samples
Blake Mason, Ardhendu Tripathy, Robert Nowak
We consider the problem of learning the nearest neighbor graph of a dataset of n items. The metric is unknown, but we can query an oracle to obtain a noisy estimate of the distance between any pair of items. This framework applies to problem domains where one wants to learn people's preferences from responses commonly modeled as noisy distance judgments. In this paper, we propose an active algorithm to find the graph with high probability and analyze its query complexity. In contrast to existing work that forces Euclidean structure, our method is valid for general metrics, assuming only symmetry and the triangle inequality. Furthermore, we demonstrate efficiency of our method empirically and theoretically, needing only O(n\log(n)\Delta^{-2}) queries in favorable settings, where \Delta^{-2} accounts for the effect of noise. Using crowd-sourced data collected for a subset of the UT~Zappos50K dataset, we apply our algorithm to learn which shoes people believe are most similar and show that it beats both an active baseline and ordinal embedding.
****
Visual Concept-Metaconcept Learning
Chi Han, Jiayuan Mao, Chuang Gan, Josh Tenenbaum, Jiajun Wu
Humans reason with concepts and metaconcepts: we recognize red and blue from visual input; we also understand that they are colors, i.e., red is an instance of color. In this paper, we propose the visual concept-metaconcept learner (VCML) for joint learning of concepts and metaconcepts from images and associated question-answer pairs. The key is to exploit the bidirectional connection between visual concepts and metaconcepts. Visual representations provide grounding cues for predicting relations between unseen pairs of concepts. Knowing that red and blue are instances of color, we generalize to the fact that green is also an instance of color since they all categorize the hue of objects. Meanwhile, knowledge about metaconcepts empowers visual concept learning from limited, noisy, and even biased data. From just a few examples of purple cubes we can understand a new color purple, which resembles the hue of the cubes instead of the shape of them. Evaluation on both synthetic and real-world datasets validates our claims.
****
The Point Where Reality Meets Fantasy: Mixed Adversarial Generators for Image Splice Detection
Vladimir V. Kniaz, Vladimir Knyaz, Fabio Remondino
Modern photo editing tools allow creating realistic manipulated images easily. While fake images can be quickly generated, learning models for their detection is challenging due to the high variety of tampering artifacts and the lack of large labeled datasets of manipulated images. In this paper, we propose a new framework for training of discriminative segmentation model via an adversarial process. We simultaneously train four models: a generative retouching model GR that translates manipulated image to the real image domain, a generative annotation model GA that estimates the pixel-wise probability of image patch being either real or fake, and two discriminators DR and DA that qualify the output of GR and GA. The aim of model GR is to maximize the probability of model GA making a mistake. Our method extends the generative adversarial networks framework with two main

contributions: (1) training of a generative model GR against a deep semantic segmentation network GA that learns rich scene semantics for manipulated region detection, (2) proposing per class semantic loss that facilitates semantically consistent image retouching by the G_R. We collected large-scale manipulated image dataset to train our model. The dataset includes 16k real and fake images with pixel-level annotations of manipulated areas. The dataset also provides ground truth pixel-level object annotations. We validate our approach on several modern manipulated image datasets, where quantitative results and ablations demonstrate that our method achieves and surpasses the state-of-the-art in manipulated image detection. We made our code and dataset publicly available.
****

Self-Supervised Deep Learning on Point Clouds by Reconstructing Space
Jonathan Sauder, Bjarne Sievers
Point clouds provide a flexible and natural representation usable in countless applications such as robotics or self-driving cars. Recently, deep neural networks operating on raw point cloud data have shown promising results on supervised learning tasks such as object classification and semantic segmentation. While massive point cloud datasets can be captured using modern scanning technology, manually labelling such large 3D point clouds for supervised learning tasks is a cumbersome process. This necessitates methods that can learn from unlabelled data to significantly reduce the number of annotated samples needed in supervised learning. We propose a self-supervised learning task for deep learning on raw point cloud data in which a neural network is trained to reconstruct point clouds whose parts have been randomly rearranged. While solving this task, representations that capture semantic properties of the point cloud are learned. Our method is agnostic of network architecture and outperforms current unsupervised learning approaches in downstream object classification tasks. We show experimentally, that pre-training with our method before supervised training improves the performance of state-of-the-art models and significantly improves sample efficiency.
****

Outlier-Robust High-Dimensional Sparse Estimation via Iterative Filtering
Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Eric Price, Alistair Stewart
We study high-dimensional sparse estimation tasks in a robust setting where a constant fraction
of the dataset is adversarially corrupted. Specifically, we focus on the fundamental problems of robust
sparse mean estimation and robust sparse PCA.
We give the first practically viable robust estimators for these problems.
In more detail, our algorithms are sample and computationally efficient
and achieve near-optimal robustness guarantees.
In contrast to prior provable algorithms which relied on the ellipsoid method,
our algorithms use spectral techniques to iteratively remove outliers from the dataset.
Our experimental evaluation on synthetic data shows that our algorithms are scalable and
significantly outperform a range of previous approaches, nearly matching the best error rate without corruptions.
****

ODE2VAE: Deep generative second order ODEs with Bayesian neural networks
Cagatay Yildiz, Markus Heinonen, Harri Lahdesmaki
We present Ordinary Differential Equation Variational Auto-Encoder (ODE2VAE), a latent second order ODE model for high-dimensional sequential data. Leveraging the advances in deep generative models, ODE2VAE can simultaneously learn the embedding of high dimensional trajectories and infer arbitrarily complex continuous-time latent dynamics. Our model explicitly decomposes the latent space into momentum and position components and solves a second order ODE system, which is in contrast to recurrent neural network (RNN) based time series models and recently proposed black-box ODE techniques. In order to account for uncertainty, we propose probabilistic latent ODE dynamics parameterized by deep Bayesian neural networks. We demonstrate our approach on motion capture, image rotation, and bouncing

balls datasets. We achieve state-of-the-art performance in long term motion prediction and imputation tasks.
****
Cross-Domain Transferability of Adversarial Perturbations
Muhammad Muzammal Naseer, Salman H. Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, Fatih Porikli
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.
****
Thresholding Bandit with Optimal Aggregate Regret
Chao Tao, Saúl Blanco, Jian Peng, 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.
****
Recovering Bandits
Ciara Pike-Burke, Steffen Grunewalder
We study the recovering bandits problem, a variant of the stochastic multi-armed bandit problem where the expected reward of each arm varies according to some unknown function of the time since the arm was last played. While being a natural extension of the classical bandit problem that arises in many real-world settings, this variation is accompanied by significant difficulties. In particular, methods need to plan ahead and estimate many more quantities than in the classical bandit setting. In this work, we explore the use of Gaussian processes to tackle the estimation and planing problem. We also discuss different regret definitions that let us quantify the performance of the methods. To improve computational efficiency of the methods, we provide an optimistic planning approximation. We complement these discussions with regret bounds and empirical studies.
****
A neurally plausible model for online recognition and postdiction in a dynamical environment
Li Kevin Wenliang, Maneesh Sahani
Humans and other animals are frequently near-optimal in their ability to integrate noisy and ambiguous sensory data to form robust percepts---which are informed both by sensory evidence and by prior expectations about the structure of the environment. It is suggested that the brain does so using the statistical structure provided by an internal model of how latent, causal factors produce the observed patterns. In dynamic environments, such integration often takes the form of \emph{postdiction}, wherein later sensory evidence affects inferences about earlier percepts. As the brain must operate in current time, without the luxury of acausal propagation of information, how does such postdictive inference come about? Here, we propose a general framework for neural probabilistic inference in dynamic models based on the distributed distributional code (DDC) representation of uncertainty, naturally extending the underlying encoding to incorporate implicit probabilistic beliefs about both present and past. We show that, as in other uses of the DDC, an inferential model can be learnt efficiently using samples from an internal model of the world. Applied to stimuli used in the context of psychophysics experiments, the framework provides an online and plausible mechanism for inference, including postdictive effects.
****
Limits of Private Learning with Access to Public Data
Noga Alon, Raef Bassily, Shay Moran
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.
****

Optimizing Generalized PageRank Methods for Seed-Expansion Community Detection
Pan Li, I Chien, Olgica Milenkovic
Landing probabilities (LP) of random walks (RW) over graphs encode rich informat ion regarding graph topology. Generalized PageRanks (GPR), which represent weigh ted sums of LPs of RWs, utilize the discriminative power of LP features to enabl e many graph-based learning studies. Previous work in the area has mostly focuse d on evaluating suitable weights for GPRs, and only a few studies so far have at tempted to derive the optimal weights of GPRs for a given application. We take a fundamental step forward in this direction by using random graph models to bett er our understanding of the behavior of GPRs. In this context, we provide a rigo rous non-asymptotic analysis for the convergence of LPs and GPRs to their mean-f ield values on edge-independent random graphs. Although our theoretical results apply to many problem settings, we focus on the task of seed-expansion community detection over stochastic block models. There, we find that the predictive powe r of LPs decreases significantly slower than previously reported based on asympt otic findings. Given this result, we propose a new GPR, termed Inverse PR (IPR), with LP weights that increase for the initial few steps of the walks. Extensive experiments on both synthetic and real, large-scale networks illustrate the sup eriority of IPR compared to other GPRs for seeded community detection.
****

Importance Resampling for Off-policy Prediction
Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, Martha White
Importance sampling (IS) is a common reweighting strategy for off-policy predict ion in reinforcement learning. While it is consistent and unbiased, it can resu lt in high variance updates to the weights for the value function. In this work, we explore a resampling strategy as an alternative to reweighting. We propose I mportance Resampling (IR) for off-policy prediction, which resamples experience from a replay buffer and applies standard on-policy updates. The approach avoids using importance sampling ratios in the update, instead correcting the distribu tion before the update. We characterize the bias and consistency of IR, particul arly compared to Weighted IS (WIS). We demonstrate in several microworlds that I R has improved sample efficiency and lower variance updates, as compared to IS a nd several variance-reduced IS strategies, including variants of WIS and V-trace which clips IS ratios. We also provide a demonstration showing IR improves over IS for learning a value function from images in a racing car simulator.
****

A Condition Number for Joint Optimization of Cycle-Consistent Networks
Leonidas J. Guibas, Qixing Huang, Zhenxiao Liang
A recent trend in optimizing maps such as dense correspondences between objects or neural networks between pairs of domains is to optimize them jointly. In this context, there is a natural \textsl{cycle-consistency} constraint, which regula rizes composite maps associated with cycles, i.e., they are forced to be identit y maps. However, as there is an exponential number of cycles in a graph, how to sample a subset of cycles becomes critical for efficient and effective enforceme nt of the cycle-consistency constraint. This paper presents an algorithm that se lect a subset of weighted cycles to minimize a condition number of the induced j oint optimization problem. Experimental results on benchmark datasets justify th e effectiveness of our approach for optimizing dense correspondences between 3D shapes and neural networks for predicting dense image flows.
****

A Graph Theoretic Additive Approximation of Optimal Transport
Nathaniel Lahn, Deepika Mulchandani, Sharath Raghvendra
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.
****

MaxGap Bandit: Adaptive Algorithms for Approximate Ranking
Sumeet Katariya, Ardhendu Tripathy, Robert Nowak
This paper studies the problem of adaptively sampling from K distributions (arms

) in order to identify the largest gap between any two adjacent means. We call this the MaxGap-bandit problem. This problem arises naturally in approximate ranking, noisy sorting, outlier detection, and top-arm identification in bandits. The key novelty of the MaxGap bandit problem is that it aims to adaptively determine the natural partitioning of the distributions into a subset with larger means and a subset with smaller means, where the split is determined by the largest gap rather than a pre-specified rank or threshold. Estimating an arm’s gap requires sampling its neighboring arms in addition to itself, and this dependence results in a novel hardness parameter that characterizes the sample complexity of the problem. We propose elimination and UCB-style algorithms and show that they are minimax optimal. Our experiments show that the UCB-style algorithms require 6-8x fewer samples than non-adaptive sampling to achieve the same error.
****
Regret Bounds for Learning State Representations in Reinforcement Learning
Ronald Ortner, Matteo Pirotta, Alessandro Lazaric, Ronan Fruit, Odalric-Ambrym Maillard
We consider the problem of online reinforcement learning when several state representations (mapping histories to a discrete state space) are available to the learning agent. At least one of these representations is assumed to induce a Markov decision process (MDP), and the performance of the agent is measured in terms of cumulative regret against the optimal policy giving the highest average reward in this MDP representation. We propose an algorithm (UCB-MS) with O(sqrt(T)) regret in any communicating Markov decision process. The regret bound shows that UCB-MS automatically adapts to the Markov model. This improves over the currently known best results in the literature that gave regret bounds of order O(T^(2/3)).
****
Exact Rate-Distortion in Autoencoders via Echo Noise
Rob Brekelmans, Daniel Moyer, Aram Galstyan, Greg Ver Steeg
Compression is at the heart of effective representation learning. However, lossy compression is typically achieved through simple parametric models like Gaussian noise to preserve analytic tractability, and the limitations this imposes on learning are largely unexplored. Further, the Gaussian prior assumptions in models such as variational autoencoders (VAEs) provide only an upper bound on the compression rate in general. We introduce a new noise channel, Echo noise, that admits a simple, exact expression for mutual information for arbitrary input distributions. The noise is constructed in a data-driven fashion that does not require restrictive distributional assumptions. With its complex encoding mechanism and exact rate regularization, Echo leads to improved bounds on log-likelihood and dominates beta-VAEs across the achievable range of rate-distortion trade-offs. Further, we show that Echo noise can outperform flow-based methods without the need to train additional distributional transformations.
****
AutoAssist: A Framework to Accelerate Training of Deep Neural Networks
Jiong Zhang, Hsiang-Fu Yu, Inderjit S. Dhillon
Deep neural networks have yielded superior performance in many contemporary
applications. However, the gradient computation in a deep model with millions of
instances leads to a lengthy training process even with modern
GPU/TPU hardware acceleration. In this paper, we propose AutoAssist, a
simple framework to accelerate training of a deep neural network.
Typically, as the training procedure evolves, the amount of improvement by a
stochastic gradient update varies dynamically with the choice of instances in the mini-batch.
In AutoAssist, we utilize this fact and design an instance shrinking operation that is
used to filter out instances with relatively low marginal improvement to the
current model; thus the computationally intensive gradient computations are
performed on informative instances as much as possible.
Specifically, we train
a very lightweight Assistant model jointly with the original deep network, which

we refer to as Boss.
The Assistant model is designed to gauge the importance of a given
instance with respect to the current Boss such that the shrinking operation can be applied in the batch generator. With careful design, we train the Boss and Assistant in a nonblocking and asynchronous fashion such that
overhead is minimal.
To demonstrate the effectiveness of AutoAssist, we conduct experiments on two contemporary applications: image classification using ResNets with varied number of layers, and neural machine translation using LSTMs, ConvS2S and
Transformer models. For each application, we verify that AutoAssist leads to significant reduction in training time; in particular, 30% to 40% of the total operation count can be reduced which leads to
faster convergence and a corresponding decrease in training time.
****
BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling
Lars Maaløe, Marco Fraccaro, Valentin Liévin, Ole Winther
With the introduction of the variational autoencoder (VAE), probabilistic latent variable models have received renewed attention as powerful generative models. However, their performance in terms of test likelihood and quality of generated samples has been surpassed by autoregressive models without stochastic units. Furthermore, flow-based models have recently been shown to be an attractive alternative that scales well to high-dimensional data. In this paper we close the performance gap by constructing VAE models that can effectively utilize a deep hierarchy of stochastic variables and model complex covariance structures. We introduce the Bidirectional-Inference Variational Autoencoder (BIVA), characterized by a skip-connected generative model and an inference network formed by a bidirectional stochastic inference path. We show that BIVA reaches state-of-the-art test likelihoods, generates sharp and coherent natural images, and uses the hierarchy of latent variables to capture different aspects of the data distribution. We observe that BIVA, in contrast to recent results, can be used for anomaly detection. We attribute this to the hierarchy of latent variables which is able to extract high-level semantic features. Finally, we extend BIVA to semi-supervised classification tasks and show that it performs comparably to state-of-the-art results by generative adversarial networks.
****
Multiway clustering via tensor block models
Miaoyan Wang, Yuchen Zeng
We consider the problem of identifying multiway block structure from a large noisy tensor. Such problems arise frequently in applications such as genomics, recommendation system, topic modeling, and sensor network localization. We propose a tensor block model, develop a unified least-square estimation, and obtain the theoretical accuracy guarantees for multiway clustering. The statistical convergence of the estimator is established, and we show that the associated clustering procedure achieves partition consistency. A sparse regularization is further developed for identifying important blocks with elevated means. The proposal handles a broad range of data types, including binary, continuous, and hybrid observations. Through simulation and application to two real datasets, we demonstrate the outperformance of our approach over previous methods.
****
Bridging Machine Learning and Logical Reasoning by Abductive Learning
Wang-Zhou Dai, Qiuling Xu, Yang Yu, Zhi-Hua Zhou
Perception and reasoning are two representative abilities of intelligence that are integrated seamlessly during human problem-solving processes. In the area of artificial intelligence (AI), the two abilities are usually realised by machine learning and logic programming, respectively. However, the two categories of techniques were developed separately throughout most of the history of AI. In this paper, we present the abductive learning targeted at unifying the two AI paradigms in a mutually beneficial way, where the machine learning model learns to perceive primitive logic facts from data, while logical reasoning can exploit symbolic domain knowledge and correct the wrongly perceived facts for improving the ma

chine learning models. Furthermore, we propose a novel approach to optimise the machine learning model and the logical reasoning model jointly. We demonstrate that by using abductive learning, machines can learn to recognise numbers and resolve unknown mathematical operations simultaneously from images of simple hand-written equations. Moreover, the learned models can be generalised to longer equations and adapted to different tasks, which is beyond the capability of state-of-the-art deep learning models.
****

Variational Structured Semantic Inference for Diverse Image Captioning
Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, Yan Wang
Despite the exciting progress in image captioning, generating diverse captions for a given image remains as an open problem. Existing methods typically apply generative models such as Variational Auto-Encoder to diversify the captions, which however neglect two key factors of diverse expression, i.e., the lexical diversity and the syntactic diversity. To model these two inherent diversities in image captioning, we propose a Variational Structured Semantic Inferring model (termed VSSI-cap) executed in a novel structured encoder-inferer-decoder schema. VSSI-cap mainly innovates in a novel structure, i.e., Variational Multi-modal Inferring tree (termed VarMI-tree). In particular, conditioned on the visual-textual features from the encoder, the VarMI-tree models the lexical and syntactic diversities by inferring their latent variables (with variations) in an approximate posterior inference guided by a visual semantic prior. Then, a reconstruction loss and the posterior-prior KL-divergence are jointly estimated to optimize the VSSI-cap model. Finally, diverse captions are generated upon the visual features and the latent variables from this structured encoder-inferer-decoder model. Experiments on the benchmark dataset show that the proposed VSSI-cap achieves significant improvements over the state-of-the-arts.
****

Input-Output Equivalence of Unitary and Contractive RNNs
Melikasadat Emami, Mojtaba Sahraee Ardakan, Sundeep Rangan, Alyson K. Fletcher
Unitary recurrent neural networks (URNNs) have been proposed as a method to overcome the vanishing and exploding gradient problem in modeling data with long-term dependencies. A basic question is how restrictive is the unitary constraint on the possible input-output mappings of such a network? This works shows that for any contractive RNN with ReLU activations, there is a URNN with at most twice the number of hidden states and the identical input-output mapping. Hence, with ReLU activations, URNNs are as expressive as general RNNs. In contrast, for certain smooth activations, it is shown that the input-output mapping of an RNN cannot be matched with a URNN, even with an arbitrary number of states. The theoretical results are supported by experiments on modeling of slowly-varying dynamical systems.
****

Latent Weights Do Not Exist: Rethinking Binarized Neural Network Optimization
Koen Helwegen, James Widdicombe, Lukas Geiger, Zechun Liu, Kwang-Ting Cheng, Roeland Nusselder
Optimization of Binarized Neural Networks (BNNs) currently relies on real-valued latent weights to accumulate small update steps. In this paper, we argue that these latent weights cannot be treated analogously to weights in real-valued networks. Instead their main role is to provide inertia during training. We interpret current methods in terms of inertia and provide novel insights into the optimization of BNNs. We subsequently introduce the first optimizer specifically designed for BNNs, Binary Optimizer (Bop), and demonstrate its performance on CIFAR-10 and ImageNet. Together, the redefinition of latent weights as inertia and the introduction of Bop enable a better understanding of BNN optimization and open up the way for further improvements in training methodologies for BNNs.
****

Nonparametric Regressive Point Processes Based on Conditional Gaussian Processes
Siqi Liu, Milos Hauskrecht
Real-world event sequences consist of complex mixtures of different types of eve

nts occurring in time. An event may depend on past events of the same type, as w
ell as, the other types. Point processes define a general class of models for ev
ent sequences. ``Regressive point processes'' refer to point processes that dire
ctly model the dependency between an event and any past event, an example of whi
ch is a Hawkes process. In this work, we propose and develop a new nonparametric
 regressive point process model based on Gaussian processes. We show that our mo
del can represent better many commonly observed real-world event sequences and c
apture the dependencies between events that are difficult to model using existin
g nonparametric Hawkes process variants. We demonstrate the improved predictive
performance of our model against state-of-the-art baselines on multiple syntheti
c and real-world datasets.
****
Continuous-time Models for Stochastic Optimization Algorithms
Antonio Orvieto, Aurelien Lucchi
We propose new continuous-time formulations for first-order stochastic optimizat
ion algorithms such as mini-batch gradient descent and variance-reduced methods.
 We exploit these continuous-time models, together with simple Lyapunov analysis
 as well as tools from stochastic calculus, in order to derive convergence bound
s for various types of non-convex functions. Guided by such analysis, we show th
at the same Lyapunov arguments hold in discrete-time, leading to matching rates.
 In addition, we use these models and Ito calculus to infer novel insights on th
e dynamics of SGD, proving that a decreasing learning rate acts as time warping
or, equivalently, as landscape stretching.
****
Differentiable Convex Optimization Layers
Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, J. Zi
co Kolter
Recent work has shown how to embed differentiable optimization problems (that is
, problems whose solutions can be backpropagated through) as layers within deep
learning architectures. This method provides a useful inductive bias for certain
 problems, but existing software for differentiable optimization layers is rigid
 and difficult to apply to new settings. In this paper, we propose an approach t
o differentiating through disciplined convex programs, a subclass of convex opti
mization problems used by domain-specific languages (DSLs) for convex optimizati
on. We introduce disciplined parametrized programming, a subset of disciplined c
onvex programming, and we show that every disciplined parametrized program can b
e represented as the composition of an affine map from parameters to problem dat
a, a solver, and an affine map from the solver’s solution to a solution of the o
riginal problem (a new form we refer to as affine-solver-affine form). We then d
emonstrate how to efficiently differentiate through each of these components, al
lowing for end-to-end analytical differentiation through the entire convex progr
am. We implement our methodology in version 1.1 of CVXPY, a popular Python-embed
ded DSL for convex optimization, and additionally implement differentiable layer
s for disciplined convex programs in PyTorch and TensorFlow 2.0. Our implementat
ion significantly lowers the barrier to using convex optimization problems in di
fferentiable programs. We present applications in linear machine learning models
 and in stochastic control, and we show that our layer is competitive (in execut
ion time) compared to specialized differentiable solvers from past work.
****
A Zero-Positive Learning Approach for Diagnosing Software Performance Regression
s
Mejbah Alam, Justin Gottschlich, Nesime Tatbul, Javier S. Turek, Tim Mattson, Ab
dullah Muzahid
The field of machine programming (MP), the automation of the development of soft
ware, is making notable research advances. This is, in part, due to the emergenc
e of a wide range of novel techniques in machine learning. In this paper, we app
ly MP to the automation of software performance regression testing. A performanc
e regression is a software performance degradation caused by a code change. We p
resent AutoPerf – a novel approach to automate regression testing that utilizes
three core techniques: (i) zero-positive learning, (ii) autoencoders, and (iii)

hardware telemetry. We demonstrate AutoPerf’s generality and efficacy against 3
types of performance regressions across 10 real performance bugs in 7 benchmark
and open-source programs. On average, AutoPerf exhibits 4% profiling
overhead and accurately diagnoses more performance bugs than prior state-of-the-
art approaches. Thus far, AutoPerf has produced no false negatives.
****
Partially Encrypted Deep Learning using Functional Encryption
Théo Ryffel, David Pointcheval, Francis Bach, Edouard Dufour-Sans, Romain Gay
Machine learning on encrypted data has received a lot of attention thanks to rec
ent breakthroughs in homomorphic encryption and secure multi-party computation.
It allows outsourcing computation to untrusted servers without sacrificing priva
cy of sensitive data. We propose a practical framework to perform partially encr
ypted and privacy-preserving predictions which combines adversarial training and
functional encryption. We first present a new functional encryption scheme to e
fficiently compute quadratic functions so that the data owner controls what can
be computed but is not involved in the calculation: it provides a decryption key
which allows one to learn a specific function evaluation of some encrypted data
. We then show how to use it in machine learning to partially encrypt neural net
works with quadratic activation functions at evaluation time and we provide a th
orough analysis of the information leaks based on indistinguishability of data i
tems of the same label. Last, since several encryption schemes cannot deal with
the last thresholding operation used for classification, we propose a training m
ethod to prevent selected sensitive features from leaking which adversarially op
timizes the network against an adversary trying to identify these features. This
is of great interest for several existing works using partially encrypted machi
ne learning as it comes with almost no cost on the model's accuracy and signific
antly improves data privacy.
****
Graph Transformer Networks
Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, Hyunwoo J. Kim
Graph neural networks (GNNs) have been widely used in representation learning on
graphs and achieved state-of-the-art performance in tasks such as node classifi
cation and link prediction. However, most existing GNNs are designed to learn no
de representations on the fixed and homogeneous graphs. The limitations especial
ly become problematic when learning representations on a misspecified graph or a
heterogeneous graph that consists of various types of nodes and edges. In this
paper, we propose Graph Transformer Networks (GTNs) that are capable of generati
ng new graph structures, which involve identifying useful connections between un
connected nodes on the original graph, while learning effective node representat
ion on the new graphs in an end-to-end fashion. Graph Transformer layer, a core
layer of GTNs, learns a soft selection of edge types and composite relations for
generating useful multi-hop connections so-call meta-paths. Our experiments sho
w that GTNs learn new graph structures, based on data and tasks without domain k
nowledge, and yield powerful node representation via convolution on the new grap
hs. Without domain-specific graph preprocessing, GTNs achieved the best performa
nce in all three benchmark node classification tasks against the state-of-the-ar
t methods that require pre-defined meta-paths from domain knowledge.
****
Non-normal Recurrent Neural Network (nnRNN): learning long time dependencies whi
le improving expressivity with transient dynamics
Giancarlo Kerg, Kyle Goyette, Maximilian Puelma Touzel, Gauthier Gidel, Eugene V
orontsov, Yoshua Bengio, Guillaume Lajoie
A recent strategy to circumvent the exploding and vanishing gradient problem in
RNNs, and to allow the stable propagation of signals over long time scales, is t
o constrain recurrent connectivity matrices to be orthogonal or unitary. This en
sures eigenvalues with unit norm and thus stable dynamics and training. However
this comes at the cost of reduced expressivity due to the limited variety of ort
hogonal transformations. We propose a novel connectivity structure based on the
Schur decomposition and a splitting of the Schur form into normal and non-normal
parts.

This allows to parametrize matrices with unit-norm eigenspectra without orthogonality constraints on eigenbases. The resulting architecture ensures access to a larger space of spectrally constrained matrices, of which orthogonal matrices are a subset.
This crucial difference retains the stability advantages and training speed of orthogonal RNNs while enhancing expressivity, especially on tasks that require computations over ongoing input sequences.
****
Large Memory Layers with Product Keys
Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, Herve Jegou
This paper introduces a structured memory which can be easily integrated into a neural network. The memory is very large by design and significantly increases the capacity of the architecture, by up to a billion parameters with a negligible computational overhead.
Its design and access pattern is based on product keys, which enable fast and exact nearest neighbor search. The ability to increase the number of parameters while keeping the same computational budget lets the overall system strike a better trade-off between prediction accuracy and computation efficiency both at training and test time. This memory layer allows us to tackle very large scale language modeling tasks. In our experiments we consider a dataset with up to 30 billion words, and we plug our memory layer in a state-of-the-art transformer-based architecture. In particular, we found that a memory augmented model with only 12 layers outperforms a baseline transformer model with 24 layers, while being twice faster at inference time. We release our code for reproducibility purposes.
****
Computing Full Conformal Prediction Set with Approximate Homotopy
Eugene Ndiaye, Ichiro Takeuchi
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.
****
AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification
Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, Shanfeng Zhu
Extreme multi-label text classification (XMTC) is an important problem in the era of {\it big data}, for tagging a given text with the most relevant multiple labels from an extremely large-scale label set. XMTC can be found in many applications, such as item categorization, web page tagging, and news annotation.
Traditionally most methods used bag-of-words (BOW) as inputs, ignoring word context as well as deep semantic information. Recent attempts to overcome the problems of BOW by deep learning still suffer from 1) failing to capture the important subtext for each label and 2) lack of scalability against the huge number of labels.
We propose a new label tree-based deep learning model for XMTC, called AttentionXML, with two unique features: 1) a multi-label attention mechanism with raw text as input, which allows to capture the most relevant part of text to each label; and 2) a shallow and wide probabilistic label tree (PLT), which allows to handle millions of labels, especially for "tail labels".
We empirically compared the performance of AttentionXML with those of eight state-of-the-art methods over six benchmark datasets, including Amazon-3M with around 3 million labels. AttentionXML outperformed all competing methods under all experimental settings.
Experimental results also show that AttentionXML achieved the best performance against tail labels among label tree-based methods. The code and datasets are available at \url{http://github.com/yourh/AttentionXML} .
****
Policy Learning for Fairness in Ranking

Ashudeep Singh, Thorsten Joachims
Conventional Learning-to-Rank (LTR) methods optimize the utility of the rankings to the users, but they are oblivious to their impact on the ranked items. However, there has been a growing understanding that the latter is important to consider for a wide range of ranking applications (e.g. online marketplaces, job placement, admissions). To address this need, we propose a general LTR framework that can optimize a wide range of utility metrics (e.g. NDCG) while satisfying fairness of exposure constraints with respect to the items. This framework expands the class of learnable ranking functions to stochastic ranking policies, which provides a language for rigorously expressing fairness specifications. Furthermore, we provide a new LTR algorithm called Fair-PG-Rank for directly searching the space of fair ranking policies via a policy-gradient approach. Beyond the theoretical evidence in deriving the framework and the algorithm, we provide empirical results on simulated and real-world datasets verifying the effectiveness of the approach in individual and group-fairness settings.
****
Regret Minimization for Reinforcement Learning by Evaluating the Optimal Bias Function
Zihan Zhang, Xiangyang Ji
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.
****
Integer Discrete Flows and Lossless Compression
Emiel Hoogeboom, Jorn Peters, Rianne van den Berg, Max Welling
Lossless compression methods shorten the expected representation size of data without loss of information, using a statistical model. Flow-based models are attractive in this setting because they admit exact likelihood optimization, which is equivalent to minimizing the expected number of bits per message. However, conventional flows assume continuous data, which may lead to reconstruction errors when quantized for compression. For that reason, we introduce a flow-based generative model for ordinal discrete data called Integer Discrete Flow (IDF): a bijective integer map that can learn rich transformations on high-dimensional data. As building blocks for IDFs, we introduce a flexible transformation layer called integer discrete coupling. Our experiments show that IDFs are competitive with other flow-based generative models. Furthermore, we demonstrate that IDF based compression achieves state-of-the-art lossless compression rates on CIFAR10, ImageNet32, and ImageNet64. To the best of our knowledge, this is the first lossless compression method that uses invertible neural networks.
****
Generative Well-intentioned Networks
Justin Cosentino, Jun Zhu
We propose Generative Well-intentioned Networks (GWINs), a novel framework for increasing the accuracy of certainty-based, closed-world classifiers. A conditional generative network recovers the distribution of observations that the classifier labels correctly with high certainty. We introduce a reject option to the classifier during inference, allowing the classifier to reject an observation instance rather than predict an uncertain label. These rejected observations are translated by the generative network to high-certainty representations, which are then relabeled by the classifier. This architecture allows for any certainty-based classifier or rejection function and is not limited to multilayer perceptrons. The capability of this framework is assessed using benchmark classification datasets and shows that GWINs significantly improve the accuracy of uncertain observations.
****
An Embedding Framework for Consistent Polyhedral Surrogates
Jessica Finocchiaro, Rafael Frongillo, Bo Waggoner
We formalize and study the natural approach of designing convex surrogate loss functions via embeddings for problems such as classification or ranking. In this

approach, one embeds each of the finitely many predictions (e.g. classes) as a p
oint in \reals^d, assigns the original loss values to these points, and convexif
ies the loss in some way to obtain a surrogate. We prove that this approach is
equivalent, in a strong sense, to working with polyhedral (piecewise linear conv
ex) losses. Moreover, given any polyhedral loss L, we give a construction of a
link function through which L is a consistent surrogate for the loss it embeds.
 We go on to illustrate the power of this embedding framework with succinct proo
fs of consistency or inconsistency of various polyhedral surrogates in the liter
ature.
****
The Normalization Method for Alleviating Pathological Sharpness in Wide Neural N
etworks
Ryo Karakida, Shotaro Akaho, Shun-ichi Amari
Normalization methods play an important role in enhancing the performance of dee
p learning while their theoretical understandings have been limited. To theoreti
cally elucidate the effectiveness of normalization, we quantify the geometry of
the parameter space determined by the Fisher information matrix (FIM), which als
o corresponds to the local shape of the loss landscape under certain conditions.
 We analyze deep neural networks with random initialization, which is known to s
uffer from a pathologically sharp shape of the landscape when the network become
s sufficiently wide. We reveal that batch normalization in the last layer contri
butes to drastically decreasing such pathological sharpness if the width and sam
ple number satisfy a specific condition. In contrast, it is hard for batch norma
lization in the middle hidden layers to alleviate pathological sharpness in many
 settings. We also found that layer normalization cannot alleviate pathological
sharpness either. Thus, we can conclude that batch normalization in the last lay
er significantly contributes to decreasing the sharpness induced by the FIM.
****
Re-randomized Densification for One Permutation Hashing and Bin-wise Consistent
Weighted Sampling
Ping Li, Xiaoyun Li, Cun-Hui Zhang
Jaccard similarity is widely used as a distance measure in many machine learning
and search applications. Typically, hashing methods are essential for the use of
Jaccard similarity to be practical in large-scale settings. For hashing binary (
0/1)
data, the idea of one permutation hashing (OPH) with densification significantly
accelerates traditional minwise hashing algorithms while providing unbiased and
accurate estimates. In this paper, we propose a strategy named “re-randomization
”
in the process of densification that could achieve the smallest variance among a
ll
densification schemes. The success of this idea naturally inspires us to general
ize
one permutation hashing to weighted (non-binary) data, which results in the soca
lled “bin-wise consistent weighted sampling (BCWS)” algorithm. We analyze the
behavior of BCWS and compare it with a recent alternative. Extensive experiments
on various datasets illustrates the effectiveness of our proposed methods.
****
Beyond Online Balanced Descent: An Optimal Algorithm for Smoothed Online Optimiz
ation
Gautam Goel, Yiheng Lin, Haoyuan Sun, Adam Wierman
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.
****
Reconciling λ-Returns with Experience Replay
Brett Daley, Christopher Amato
Modern deep reinforcement learning methods have departed from the incremental le
arning required for eligibility traces, rendering the implementation of the λ-re

turn difficult in this context. In particular, off-policy methods that utilize e xperience replay remain problematic because their random sampling of minibatches is not conducive to the efficient calculation of λ-returns. Yet replay-based me thods are often the most sample efficient, and incorporating λ-returns into them is a viable way to achieve new state-of-the-art performance. Towards this, we p ropose the first method to enable practical use of λ-returns in arbitrary replay -based methods without relying on other forms of decorrelation such as asynchron ous gradient updates. By promoting short sequences of past transitions into a sm all cache within the replay memory, adjacent λ-returns can be efficiently precom puted by sharing Q-values. Computation is not wasted on experiences that are nev er sampled, and stored λ-returns behave as stable temporal-difference (TD) targe ts that replace the target network. Additionally, our method grants the unique a bility to observe TD errors prior to sampling; for the first time, transitions c an be prioritized by their true significance rather than by a proxy to it. Furth ermore, we propose the novel use of the TD error to dynamically select λ-values that facilitate faster learning. We show that these innovations can enhance the performance of DQN when playing Atari 2600 games, even under partial observabili ty. While our work specifically focuses on λ-returns, these ideas are applicable to any multi-step return estimator.
****

Sinkhorn Barycenters with Free Support via Frank-Wolfe Algorithm

Giulia Luise, Saverio Salzo, Massimiliano Pontil, Carlo Ciliberto

We present a novel algorithm to estimate the barycenter of arbitrary probability distributions with respect to the Sinkhorn divergence. Based on a Frank-Wolfe o ptimization strategy, our approach proceeds by populating the support of the bar ycenter incrementally, without requiring any pre-allocation. We consider discret e as well as continuous distributions, proving convergence rates of the proposed algorithm in both settings. Key elements of our analysis are a new result showi ng that the Sinkhorn divergence on compact domains has Lipschitz continuous grad ient with respect to the Total Variation and a characterization of the sample co mplexity of Sinkhorn potentials. Experiments validate the effectiveness of our m ethod in practice.
****

Finite-Sample Analysis for SARSA with Linear Function Approximation

Shaofeng Zou, Tengyu Xu, Yingbin Liang

SARSA is an on-policy algorithm to learn a Markov decision process policy in rei nforcement learning. We investigate the SARSA algorithm with linear function app roximation under the non-i.i.d.\ setting, where a single sample trajectory is av ailable. With a Lipschitz continuous policy improvement operator that is smooth enough, SARSA has been shown to converge asymptotically. However, its non-asympt otic analysis is challenging and remains unsolved due to the non-i.i.d. samples , and the fact that the behavior policy changes dynamically with time. In this p aper, we develop a novel technique to explicitly characterize the stochastic bia s of a type of stochastic approximation procedures with time-varying Markov tran sition kernels. Our approach enables non-asymptotic convergence analyses of this type of stochastic approximation algorithms, which may be of independent intere st. Using our bias characterization technique and a gradient descent type of a nalysis, we further provide the finite-sample analysis on the mean square error of the SARSA algorithm. In the end, we present a fitted SARSA algorithm, whic h includes the original SARSA algorithm and its variant as special cases. This f itted SARSA algorithm provides a framework for \textit{iterative} on-policy fitt ed policy iteration, which is more memory and computationally efficient. For thi s fitted SARSA algorithm, we also present its finite-sample analysis.
****

Aligning Visual Regions and Textual Concepts for Semantic-Grounded Image Represe ntations

Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, Xu Sun

In vision-and-language grounding problems, fine-grained representations of the i mage are considered to be of paramount importance. Most of the current systems i ncorporate visual features and textual concepts as a sketch of an image. However

, plainly inferred representations are usually undesirable in that they are composed of separate components, the relations of which are elusive. In this work, we aim at representing an image with a set of integrated visual regions and corresponding textual concepts, reflecting certain semantics. To this end, we build the Mutual Iterative Attention (MIA) module, which integrates correlated visual features and textual concepts, respectively, by aligning the two modalities. We evaluate the proposed approach on two representative vision-and-language grounding tasks, i.e., image captioning and visual question answering. In both tasks, the semantic-grounded image representations consistently boost the performance of the baseline models under all metrics across the board. The results demonstrate that our approach is effective and generalizes well to a wide range of models for image-related applications. (The code is available at \url{https://github.com/fenglinliu98/MIA)
****
Network Pruning via Transformable Architecture Search
Xuanyi Dong, Yi Yang
Network pruning reduces the computation costs of an over-parameterized network without performance damage. Prevailing pruning algorithms pre-define the width and depth of the pruned networks, and then transfer parameters from the unpruned network to pruned networks. To break the structure limitation of the pruned networks, we propose to apply neural architecture search to search directly for a network with flexible channel and layer sizes. The number of the channels/layers is learned by minimizing the loss of the pruned networks. The feature map of the pruned network is an aggregation of K feature map fragments (generated by K networks of different sizes), which are sampled based on the probability distribution. The loss can be back-propagated not only to the network weights, but also to the parameterized distribution to explicitly tune the size of the channels/layers. Specifically, we apply channel-wise interpolation to keep the feature map with different channel sizes aligned in the aggregation procedure. The maximum probability for the size in each distribution serves as the width and depth of the pruned network, whose parameters are learned by knowledge transfer, e.g., knowledge distillation, from the original networks. Experiments on CIFAR-10, CIFAR-100 and ImageNet demonstrate the effectiveness of our new perspective of network pruning compared to traditional network pruning algorithms. Various searching and knowledge transfer approaches are conducted to show the effectiveness of the two components. Code is at: https://github.com/D-X-Y/NAS-Projects
****
Regret Minimization for Reinforcement Learning with Vectorial Feedback and Complex Objectives
Wang Chi Cheung
We consider an agent who is involved in an online Markov decision process, and receives a vector of outcomes every round. The agent aims to simultaneously optimize multiple objectives associated with the multi-dimensional outcomes. Due to state transitions, it is challenging to balance the vectorial outcomes for achieving near-optimality. In particular, contrary to the single objective case, stationary policies are generally sub-optimal. We propose a no-regret algorithm based on the Frank-Wolfe algorithm (Frank and Wolfe 1956), UCRL2 (Jaksch et al. 2010), as well as a crucial and novel gradient threshold procedure. The procedure involves carefully delaying gradient updates, and returns a non-stationary policy that diversifies the outcomes for optimizing the objectives.
****
Online Stochastic Shortest Path with Bandit Feedback and Unknown Transition Function
Aviv Rosenberg, Yishay Mansour
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.
****
Selective Sampling-based Scalable Sparse Subspace Clustering

Shin Matsushima, Maria Brbic
Sparse subspace clustering (SSC) represents each data point as a sparse linear combination of other data points in the dataset. In the representation learning step SSC finds a lower dimensional representation of data points, while in the spectral clustering step data points are clustered according to the underlying subspaces. However, both steps suffer from high computational and memory complexity, preventing the application of SSC to large-scale datasets. To overcome this limitation, we introduce Selective Sampling-based Scalable Sparse Subspace Clustering (S5C) algorithm which selects subsamples based on the approximated subgradients and linearly scales with the number of data points in terms of time and memory requirements. Along with the computational advantages, we derive theoretical guarantees for the correctness of S5C. Our theoretical result presents novel contribution for SSC in the case of limited number of subsamples. Extensive experimental results demonstrate effectiveness of our approach.
****
On the Expressive Power of Deep Polynomial Neural Networks
Joe Kileel, Matthew Trager, Joan Bruna
We study deep neural networks with polynomial activations, particularly their expressive power. For a fixed architecture and activation degree, a polynomial neural network defines an algebraic map from weights to polynomials. The image of this map is the functional space associated to the network, and it is an irreducible algebraic variety upon taking closure. This paper proposes the dimension of this variety as a precise measure of the expressive power of polynomial neural networks. We obtain several theoretical results regarding this dimension as a function of architecture, including an exact formula for high activation degrees, as well as upper and lower bounds on layer widths in order for deep polynomials networks to fill the ambient functional space. We also present computational evidence that it is profitable in terms of expressiveness for layer widths to increase monotonically and then decrease monotonically. Finally, we link our study to favorable optimization properties when training weights, and we draw intriguing connections with tensor and polynomial decompositions.
****
BehaveNet: nonlinear embedding and Bayesian neural decoding of behavioral videos
Eleanor Batty, Matthew Whiteway, Shreya Saxena, Dan Biderman, Taiga Abe, Simon Musall, Winthrop Gillis, Jeffrey Markowitz, Anne Churchland, John P. Cunningham, Sandeep R. Datta, Scott Linderman, Liam Paninski
A fundamental goal of systems neuroscience is to understand the relationship between neural activity and behavior. Behavior has traditionally been characterized by low-dimensional, task-related variables such as movement speed or response times. More recently, there has been a growing interest in automated analysis of high-dimensional video data collected during experiments. Here we introduce a probabilistic framework for the analysis of behavioral video and neural activity. This framework provides tools for compression, segmentation, generation, and decoding of behavioral videos. Compression is performed using a convolutional autoencoder (CAE), which yields a low-dimensional continuous representation of behavior. We then use an autoregressive hidden Markov model (ARHMM) to segment the CAE representation into discrete "behavioral syllables." The resulting generative model can be used to simulate behavioral video data. Finally, based on this generative model, we develop a novel Bayesian decoding approach that takes in neural activity and outputs probabilistic estimates of the full-resolution behavioral video. We demonstrate this framework on two different experimental paradigms using distinct behavioral and neural recording technologies.
****
Accurate Layerwise Interpretable Competence Estimation
Vickram Rajendran, William LeVine
Estimating machine learning performance “in the wild” is both an important and unsolved problem. In this paper, we seek to examine, understand, and predict the pointwise competence of classification models. Our contributions are twofold: First, we establish a statistically rigorous definition of competence that generalizes

the common notion of classifier confidence; second, we present the ALICE (Accurate Layerwise Interpretable Competence Estimation) Score, a pointwise competence estimator for any classifier. By considering distributional, data, an
d
model uncertainty, ALICE empirically shows accurate competence estimation in common failure situations such as class-imbalanced datasets, out-of-distribution datasets, and poorly trained models.
****
On the Global Convergence of (Fast) Incremental Expectation Maximization Methods
Belhal Karimi, Hoi-To Wai, Eric Moulines, Marc Lavielle
The EM algorithm is one of the most popular algorithm for inference in latent da ta models. The original formulation of the EM algorithm does not scale to large data set, because the whole data set is required at each iteration of the algori thm. To alleviate this problem, Neal and Hinton [1998] have proposed an incremen tal version of the EM (iEM) in which at each iteration the conditional expectati on of the latent data (E-step) is updated only for a mini-batch of observations. Another approach has been proposed by Cappe and Moulines [2009] in which the E-step is replaced by a stochastic approximation step, closely related to stochast ic gradient. In this paper, we analyze incremental and stochastic version of the EM algorithm as well as the variance reduced-version of [Chen et al., 2018] in a common unifying framework. We also introduce a new version incremental version , inspired by the SAGA algorithm by Defazio et al. [2014]. We establish non-asym ptotic convergence bounds for global convergence. Numerical applications are pre sented in this article to illustrate our findings.
****
Nonconvex Low-Rank Tensor Completion from Noisy Data
Changxiao Cai, Gen Li, H. Vincent Poor, Yuxin 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Gossip-based Actor-Learner Architectures for Deep Reinforcement Learning
Mahmoud Assran, Joshua Romoff, Nicolas Ballas, Joelle Pineau, Michael Rabbat
Multi-simulator training has contributed to the recent success of Deep Reinforce ment Learning (Deep RL) by stabilizing learning and allowing for higher training throughputs. In this work, we propose Gossip-based Actor-Learner Architectures (GALA) where several actor-learners (such as A2C agents) are organized in a peer -to-peer communication topology, and exchange information through asynchronous g ossip in order to take advantage of a large number of distributed simulators. We prove that GALA agents remain within an epsilon-ball of one-another during trai ning when using loosely coupled asynchronous communication. By reducing the amou nt of synchronization between agents, GALA is more computationally efficient and scalable compared to A2C, its fully-synchronous counterpart. GALA also outperfo rms A2C, being more robust and sample efficient. We show that we can run several loosely coupled GALA agents in parallel on a single GPU and achieve significant ly higher hardware utilization and frame-rates than vanilla A2C at comparable po wer draws.
****
Fast and Accurate Stochastic Gradient Estimation
Beidi Chen, Yingchen Xu, Anshumali Shrivastava
Stochastic Gradient Descent or SGD is the most popular optimization algorithm fo r large-scale problems. SGD estimates the gradient by uniform sampling with samp le size one. There have been several other works that suggest faster epoch-wise convergence by using weighted non-uniform sampling for better gradient estimates . Unfortunately, the per-iteration cost of maintaining this adaptive distributio n for gradient estimation is more than calculating the full gradient itself, whi ch we call the chicken-and-the-egg loop. As a result, the false impression of fa ster convergence in iterations, in reality, leads to slower convergence in time. In this paper, we break this barrier by providing the first demonstration of a

scheme, Locality sensitive hashing (LSH) sampled Stochastic Gradient Descent (LGD), which leads to superior gradient estimation while keeping the sampling cost per iteration similar to that of the uniform sampling. Such an algorithm is possible due to the sampling view of LSH, which came to light recently. As a consequence of superior and fast estimation, we reduce the running time of all existing gradient descent algorithms, that relies on gradient estimates including Adam, Ada-grad, etc. We demonstrate the effectiveness of our proposal with experiments on linear models as well as the non-linear BERT, which is a recent popular deep learning based language representation model.
****

Learning Disentangled Representations for Recommendation
Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, Wenwu Zhu
User behavior data in recommender systems are driven by the complex interactions of many latent factors behind the users’ decision making processes. The factors are highly entangled, and may range from high-level ones that govern user intentions, to low-level ones that characterize a user’s preference when executing an intention. Learning representations that uncover and disentangle these latent factors can bring enhanced robustness, interpretability, and controllability. However, learning such disentangled representations from user behavior is challenging, and remains largely neglected by the existing literature. In this paper, we present the MACRo-mIcro Disentangled Variational Auto-Encoder (MacridVAE) for learning disentangled representations from user behavior. Our approach achieves macro disentanglement by inferring the high-level concepts associated with user intentions (e.g., to buy a shirt or a cellphone), while capturing the preference of a user regarding the different concepts separately. A micro-disentanglement regularizer, stemming from an information-theoretic interpretation of VAEs, then forces each dimension of the representations to independently reflect an isolated low-level factor (e.g., the size or the color of a shirt). Empirical results show that our approach can achieve substantial improvement over the state-of-the-art baselines. We further demonstrate that the learned representations are interpretable and controllable, which can potentially lead to a new paradigm for recommendation where users are given fine-grained control over targeted aspects of the recommendation lists.
****

Learning Latent Process from High-Dimensional Event Sequences via Efficient Sampling
Qitian Wu, Zixuan Zhang, Xiaofeng Gao, Junchi Yan, Guihai Chen
We target modeling latent dynamics in high-dimension marked event sequences without any prior knowledge about marker relations. Such problem has been rarely studied by previous works which would have fundamental difficulty to handle the arisen challenges: 1) the high-dimensional markers and unknown relation network among them pose intractable obstacles for modeling the latent dynamic process; 2) one observed event sequence may concurrently contain several different chains of interdependent events; 3) it is hard to well define the distance between two high-dimension event sequences. To these ends, in this paper, we propose a seminal adversarial imitation learning framework for high-dimension event sequence generation which could be decomposed into: 1) a latent structural intensity model that estimates the adjacent nodes without explicit networks and learns to capture the temporal dynamics in the latent space of markers over observed sequence; 2) an efficient random walk based generation model that aims at imitating the generation process of high-dimension event sequences from a bottom-up view; 3) a discriminator specified as a seq2seq network optimizing the rewards to help the generator output event sequences as real as possible. Experimental results on both synthetic and real-world datasets demonstrate that the proposed method could effectively detect the hidden network among markers and make decent prediction for future marked events, even when the number of markers scales to million level.
****

Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty
Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, Dawn Song
Self-supervision provides effective representations for downstream tasks without

requiring labels. However, existing approaches lag behind fully supervised training and are often not thought beneficial beyond obviating or reducing the need for annotations. We find that self-supervision can benefit robustness in a variety of ways, including robustness to adversarial examples, label corruption, and common input corruptions. Additionally, self-supervision greatly benefits out-of-distribution detection on difficult, near-distribution outliers, so much so that it exceeds the performance of fully supervised methods. These results demonstrate the promise of self-supervision for improving robustness and uncertainty estimation and establish these tasks as new axes of evaluation for future self-supervised learning research.
****
Space and Time Efficient Kernel Density Estimation in High Dimensions
Arturs Backurs, Piotr Indyk, Tal Wagner
Recently, Charikar and Siminelakis (2017) presented a framework for kernel density estimation in provably sublinear query time, for kernels that possess a certain hashing-based property. However, their data structure requires a significantly increased super-linear storage space, as well as super-linear preprocessing time. These limitations inhibit the practical applicability of their approach on large datasets.
****
Random Projections with Asymmetric Quantization
Xiaoyun Li, Ping Li
The method of random projection has been a popular tool for data compression, similarity search, and machine learning. In many practical scenarios, applying quantization on randomly projected data could be very helpful to further reduce storage cost and facilitate more efficient retrievals, while only suffering from little loss in accuracy. In real-world applications, however, data collected from
different sources may be quantized under different schemes, which calls for a need to study the asymmetric quantization problem. In this paper, we investigate the cosine similarity estimators derived in such setting under the Lloyd-Max (LM) quantization scheme. We thoroughly analyze the biases and variances of a series of estimators including the basic simple estimators, their normalized versions, and
their debiased versions. Furthermore, by studying the monotonicity, we show that the expectation of proposed estimators increases with the true cosine similarity,
on a broader family of stair-shaped quantizers. Experiments on nearest neighbor search justify the theory and illustrate the effectiveness of our proposed estimators.
****
Scalable Deep Generative Relational Model with High-Order Node Dependence
Xuhui Fan, Bin Li, Caoyuan Li, Scott SIsson, Ling Chen
In this work, we propose a probabilistic framework for relational data modelling and latent structure exploring. Given the possible feature information for the nodes in a network, our model builds up a deep architecture that can approximate to the possible nonlinear mappings between the nodes' feature information and latent representations. For each node, we incorporate all its neighborhoods' high-order structure information to generate latent representation, such that these latent representations are ``smooth'' in terms of the network. Since the latent representations are generated from Dirichlet distributions, we further develop a data augmentation trick to enable efficient Gibbs sampling for Ber-Poisson likelihood with Dirichlet random variables. Our model can be ready to apply to large sparse network as its computations cost scales to the number of positive links in the networks. The superior performance of our model is demonstrated through improved link prediction performance on a range of real-world datasets.
****
Better Exploration with Optimistic Actor Critic
Kamil Ciosek, Quan Vuong, Robert Loftin, Katja Hofmann
Actor-critic methods, a type of model-free Reinforcement Learning, have been suc

cessfully applied to challenging tasks in continuous control, often achieving state-of-the art performance. However, wide-scale adoption of these methods in real-world domains is made difficult by their poor sample efficiency. We address this problem both theoretically and empirically. On the theoretical side, we identify two phenomena preventing efficient exploration in existing state-of-the-art algorithms such as Soft Actor Critic. First, combining a greedy actor update with a pessimistic estimate of the critic leads to the avoidance of actions that the agent does not know about, a phenomenon we call pessimistic underexploration. Second, current algorithms are directionally uninformed, sampling actions with equal probability in opposite directions from the current mean. This is wasteful, since we typically need actions taken along certain directions much more than others. To address both of these phenomena, we introduce a new algorithm, Optimistic Actor Critic, which approximates a lower and upper confidence bound on the state-action value function. This allows us to apply the principle of optimism in the face of uncertainty to perform directed exploration using the upper bound while still using the lower bound to avoid overestimation. We evaluate OAC in several challenging continuous control tasks, achieving state-of the art sample efficiency.
****
Algorithmic Analysis and Statistical Estimation of SLOPE via Approximate Message Passing
Zhiqi Bu, Jason Klusowski, Cynthia Rush, Weijie Su
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-objects Generation with Amortized Structural Regularization
Taufik Xu, Chongxuan LI, Jun Zhu, Bo Zhang
Deep generative models (DGMs) have shown promise in image generation. However, most of the existing methods learn a model by simply optimizing a divergence between the marginal distributions of the model and the data, and often fail to capture rich structures, such as attributes of objects and their relationships, in an image.
Human knowledge is a crucial element to the success of DGMs to infer these structures, especially in unsupervised learning.
In this paper, we propose amortized structural regularization (ASR), which adopts posterior regularization (PR) to embed human knowledge into DGMs via a set of structural constraints.
We derive a lower bound of the regularized log-likelihood in PR and adopt the amortized inference technique to jointly optimize the generative model and an auxiliary recognition model for inference efficiently.
Empirical results show that ASR outperforms the DGM baselines in terms of inference performance and sample quality.
****
A Family of Robust Stochastic Operators for Reinforcement Learning
Yingdong Lu, Mark Squillante, Chai Wah Wu
We consider a new family of stochastic operators for reinforcement learning with the goal of alleviating negative effects and becoming more robust to approximation or estimation errors. Various theoretical results are established, which include showing that our family of operators preserve optimality and increase the action gap in a stochastic sense. Our empirical results illustrate the strong benefits of our robust stochastic operators, significantly outperforming the classical Bellman operator and recently proposed operators.
****
One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers
Ari Morcos, Haonan Yu, Michela Paganini, Yuandong Tian
The success of lottery ticket initializations (Frankle and Carbin, 2019) suggests that small, sparsified networks can be trained so long as the network is initi

alized appropriately. Unfortunately, finding these "winning ticket'' initializations is computationally expensive. One potential solution is to reuse the same winning tickets across a variety of datasets and optimizers. However, the generality of winning ticket initializations remains unclear. Here, we attempt to answer this question by generating winning tickets for one training configuration (optimizer and dataset) and evaluating their performance on another configuration. Perhaps surprisingly, we found that, within the natural images domain, winning ticket initializations generalized across a variety of datasets, including Fashion MNIST, SVHN, CIFAR-10/100, ImageNet, and Places365, often achieving performance close to that of winning tickets generated on the same dataset. Moreover, winning tickets generated using larger datasets consistently transferred better than those generated using smaller datasets. We also found that winning ticket initializations generalize across optimizers with high performance. These results suggest that winning ticket initializations generated by sufficiently large datasets contain inductive biases generic to neural networks more broadly which improve training across many settings and provide hope for the development of better initialization methods.
****
Learning Distributions Generated by One-Layer ReLU Networks
Shanshan Wu, Alexandros G. Dimakis, Sujay Sanghavi
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.
****
Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules
Niklas Gebauer, Michael Gastegger, Kristof Schütt
Deep learning has proven to yield fast and accurate predictions of quantum-chemical properties to accelerate the discovery of novel molecules and materials. As an exhaustive exploration of the vast chemical space is still infeasible, we require generative models that guide our search towards systems with desired properties. While graph-based models have previously been proposed, they are restricted by a lack of spatial information such that they are unable to recognize spatial isomerism and non-bonded interactions. Here, we introduce a generative neural network for 3d point sets that respects the rotational invariance of the targeted structures. We apply it to the generation of molecules and demonstrate its ability to approximate the distribution of equilibrium structures using spatial metrics as well as established measures from chemoinformatics. As our model is able to capture the complex relationship between 3d geometry and electronic properties, we bias the distribution of the generator towards molecules with a small HOMO-LUMO gap - an important property for the design of organic solar cells.
****
Quantum Entropy Scoring for Fast Robust Mean Estimation and Improved Outlier Detection
Yihe Dong, Samuel Hopkins, 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.
****
Semi-flat minima and saddle points by embedding neural networks to overparameterization
Kenji Fukumizu, Shoichiro Yamaguchi, Yoh-ichi Mototake, Mirai Tanaka
We theoretically study the landscape of the training error for neural networks in overparameterized cases. We consider three basic methods for embedding a network into a wider one with more hidden units, and discuss whether a minimum point of the narrower network gives a minimum or saddle point of the wider one. Our results show that the networks with smooth and ReLU activation have different partially flat landscapes around the embedded point. We also relate these results

to a difference of their generalization abilities in overparameterized realization.
****
Privacy-Preserving Classification of Personal Text Messages with Secure Multi-Party Computation
Devin Reich, Ariel Todoki, Rafael Dowsley, Martine De Cock, anderson nascimento
Classification of personal text messages has many useful applications in surveillance, e-commerce, and mental health care, to name a few. Giving applications access to personal texts can easily lead to (un)intentional privacy violations. We propose the first privacy-preserving solution for text classification that is provably secure. Our method, which is based on Secure Multiparty Computation (SMC), encompasses both feature extraction from texts, and subsequent classification with logistic regression and tree ensembles. We prove that when using our secure text classification method, the application does not learn anything about the text, and the author of the text does not learn anything about the text classification model used by the application beyond what is given by the classification result itself. We perform end-to-end experiments with an application for detecting hate speech against women and immigrants, demonstrating excellent runtime results without loss of accuracy.
****
Locally Private Gaussian Estimation
Matthew Joseph, Janardhan Kulkarni, Jieming Mao, Steven Z. Wu
We study a basic private estimation problem: each of n users draws a single i.i.d. sample from an unknown Gaussian distribution N(\mu,\sigma^2), and the goal is to estimate \mu while guaranteeing local differential privacy for each user. As minimizing the number of rounds of interaction is important in the local setting, we provide adaptive two-round solutions and nonadaptive one-round solutions to this problem. We match these upper bounds with an information-theoretic lower bound showing that our accuracy guarantees are tight up to logarithmic factors for all sequentially interactive locally private protocols.
****
Distributed Low-rank Matrix Factorization With Exact Consensus
Zhihui Zhu, Qiuwei Li, Xinshuo Yang, Gongguo Tang, Michael B. Wakin
Low-rank matrix factorization is a problem of broad importance, owing to the ubiquity of low-rank models in machine learning contexts. In spite of its non- convexity, this problem has a well-behaved geometric landscape, permitting local search algorithms such as gradient descent to converge to global minimizers. In this paper, we study low-rank matrix factorization in the distributed setting, where local variables at each node encode parts of the overall matrix factors, and consensus is encouraged among certain such variables. We identify conditions under which this new problem also has a well-behaved geometric landscape, and we propose an extension of distributed gradient descent (DGD) to solve this problem. The favorable landscape allows us to prove convergence to global optimality with exact consensus, a stronger result than what is provided by off-the-shelf DGD theory.
****
Tensor Monte Carlo: Particle Methods for the GPU era
Laurence Aitchison
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 Mixtures of Plackett-Luce Models from Structured Partial Orders
Zhibing Zhao, Lirong Xia
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.
****

Combining Generative and Discriminative Models for Hybrid Inference
Victor Garcia Satorras, Zeynep Akata, Max Welling
A graphical model is a structured representation of the data generating process. The traditional method to reason over random variables is to perform inference in this graphical model. However, in many cases the generating process is only a poor approximation of the much more complex true data generating process, leading to suboptimal estimation. The subtleties of the generative process are however captured in the data itself and we can ``learn to infer'', that is, learn a direct mapping from observations to explanatory latent variables. In this work we propose a hybrid model that combines graphical inference with a learned inverse model, which we structure as in a graph neural network, while the iterative algorithm as a whole is formulated as a recurrent neural network. By using cross-validation we can automatically balance the amount of work performed by graphical inference versus learned inference. We apply our ideas to the Kalman filter, a Gaussian hidden Markov model for time sequences, and show, among other things, that our model can estimate the trajectory of a noisy chaotic Lorenz Attractor much more accurately than either the learned or graphical inference run in isolation.
****
Trust Region-Guided Proximal Policy Optimization
Yuhui Wang, Hao He, Xiaoyang Tan, Yaozhong Gan
Proximal policy optimization (PPO) is one of the most popular deep reinforcement learning (RL) methods, achieving state-of-the-art performance across a wide range of challenging tasks. However, as a model-free RL method, the success of PPO relies heavily on the effectiveness of its exploratory policy search. In this paper, we give an in-depth analysis on the exploration behavior of PPO, and show that PPO is prone to suffer from the risk of lack of exploration especially under the case of bad initialization, which may lead to the failure of training or being trapped in bad local optima. To address these issues, we proposed a novel policy optimization method, named Trust Region-Guided PPO (TRGPPO), which adaptively adjusts the clipping range within the trust region. We formally show that this method not only improves the exploration ability within the trust region but enjoys a better performance bound compared to the original PPO as well. Extensive experiments verify the advantage of the proposed method.
****
Region Mutual Information Loss for Semantic Segmentation
Shuai Zhao, Yang Wang, Zheng Yang, Deng Cai
Semantic segmentation is a fundamental problem in computer vision.
It is considered as a pixel-wise classification problem in practice,
and most segmentation models use a pixel-wise loss as their optimization criterion.
However, the pixel-wise loss ignores the dependencies between pixels in an image.
Several ways to exploit the relationship between pixels have been investigated,
\eg, conditional random fields (CRF) and pixel affinity based methods.
Nevertheless, these methods usually require additional model
branches, large extra memories, or more inference time.
In this paper, we develop a region mutual information (RMI) loss
to model the dependencies among pixels more simply and efficiently.
In contrast to the pixel-wise loss which treats the pixels as independent samples,
RMI uses one pixel and its neighbour pixels to represent this pixel.
Then for each pixel in an image,
we get a multi-dimensional point that encodes the relationship between pixels,
and the image is cast into a multi-dimensional distribution of
these high-dimensional points.
The prediction and ground truth thus can achieve high order consistency
through maximizing the mutual information (MI) between their multi-dimension

al distributions.
    Moreover, as the actual value of the MI is hard to calculate,
    we derive a lower bound of the MI and maximize the lower bound to maximize the
    real value of the MI.
    RMI only requires a few extra computational resources in the training stage,
    and there is no overhead during testing.
    Experimental results demonstrate
    that RMI can achieve substantial and consistent improvements in performance on PASCAL VOC 2012 and CamVid datasets.
    The code is available at \url{https://github.com/ZJULearning/RMI}.
****
A Stochastic Composite Gradient Method with Incremental Variance Reduction
Junyu Zhang, Lin Xiao
We consider the problem of minimizing the composition of a smooth (nonconvex) function and a smooth vector mapping, where the inner mapping is in the form of an expectation over some random variable or a finite sum. We propose a stochastic composite gradient method that employs incremental variance-reduced estimators for both the inner vector mapping and its Jacobian. We show that this method achieves the same orders of complexity as the best known first-order methods for minimizing expected-value and finite-sum nonconvex functions, despite the additional outer composition which renders the composite gradient estimator biased. This finding enables a much broader range of applications in machine learning to benefit from the low complexity of incremental variance-reduction methods.
****
An adaptive nearest neighbor rule for classification
Akshay Balsubramani, Sanjoy Dasgupta, yoav Freund, Shay Moran
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.
****
Variational Graph Recurrent Neural Networks
Ehsan Hajiramezanali, Arman Hasanzadeh, Krishna Narayanan, Nick Duffield, Mingyuan Zhou, Xiaoning Qian
Representation learning over graph structured data has been mostly studied in static graph settings while efforts for modeling dynamic graphs are still scant. In this paper, we develop a novel hierarchical variational model that introduces additional latent random variables to jointly model the hidden states of a graph recurrent neural network (GRNN) to capture both topology and node attribute changes in dynamic graphs. We argue that the use of high-level latent random variables in this variational GRNN (VGRNN) can better capture potential variability observed in dynamic graphs as well as the uncertainty of node latent representation. With semi-implicit variational inference developed for this new VGRNN architecture (SI-VGRNN), we show that flexible non-Gaussian latent representations can further help dynamic graph analytic tasks. Our experiments with multiple real-world dynamic graph datasets demonstrate that SI-VGRNN and VGRNN consistently outperform the existing baseline and state-of-the-art methods by a significant margin in dynamic link prediction.
****
Stochastic Bandits with Context Distributions
Johannes Kirschner, Andreas Krause
We introduce a stochastic contextual bandit model where at each time step the environment chooses a distribution over a context set and samples the context from this distribution. The learner observes only the context distribution while the exact context realization remains hidden. This allows for a broad range of applications where the context is stochastic or when the learner needs to predict the context. We leverage the UCB algorithm to this setting and show that it achieves an order-optimal high-probability bound on the cumulative regret for linear and kernelized reward functions. Our results strictly generalize previous work in

the sense that both our model and the algorithm reduce to the standard setting when the environment chooses only Dirac delta distributions and therefore provides the exact context to the learner. We further analyze a variant where the learner observes the realized context after choosing the action. Finally, we demonstrate the proposed method on synthetic and real-world datasets.
****
Geometry-Aware Neural Rendering
Joshua Tobin, Wojciech Zaremba, Pieter Abbeel
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.
****
Training Language GANs from Scratch
Cyprien de Masson d'Autume, Shakir Mohamed, Mihaela Rosca, Jack Rae
Generative Adversarial Networks (GANs) enjoy great success at image generation, but have proven difficult to train in the domain of natural language. Challenges with gradient estimation, optimization instability, and mode collapse have lead practitioners to resort to maximum likelihood pre-training, followed by small amounts of adversarial fine-tuning. The benefits of GAN fine-tuning for language generation are unclear, as the resulting models produce comparable or worse samples than traditional language models. We show it is in fact possible to train a language GAN from scratch --- without maximum likelihood pre-training. We combine existing techniques such as large batch sizes, dense rewards and discriminator regularization to stabilize and improve language GANs. The resulting model, ScratchGAN, performs comparably to maximum likelihood training on EMNLP2017 News and WikiText-103 corpora
according to quality and diversity metrics.
****
Generalization Bounds in the Predict-then-Optimize Framework
Othman El Balghiti, Adam N. Elmachtoub, Paul Grigas, Ambuj Tewari
The predict-then-optimize framework is fundamental in many practical settings: predict the unknown parameters of an optimization problem, and then solve the problem using the predicted values of the parameters. A natural loss function in this environment is to consider the cost of the decisions induced by the predicted parameters, in contrast to the prediction error of the parameters. This loss function was recently introduced in [Elmachtoub and Grigas, 2017], which called it the Smart Predict-then-Optimize (SPO) loss. Since the SPO loss is nonconvex and noncontinuous, standard results for deriving generalization bounds do not apply. In this work, we provide an assortment of generalization bounds for the SPO loss function. In particular, we derive bounds based on the Natarajan dimension that, in the case of a polyhedral feasible region, scale at most logarithmically in the number of extreme points, but, in the case of a general convex set, have poor dependence on the dimension. By exploiting the structure of the SPO loss function and an additional strong convexity assumption on the feasible region, we can dramatically improve the dependence on the dimension via an analysis and corresponding bounds that are akin to the margin guarantees in classification problems.
****
Almost Horizon-Free Structure-Aware Best Policy Identification with a Generative Model
Andrea Zanette, Mykel J. Kochenderfer, Emma Brunskill
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 the (In)fidelity and Sensitivity of Explanations
Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, Pradeep K. Ravikumar

We consider objective evaluation measures of saliency explanations for complex black-box machine learning models. We propose simple robust variants of two notions that have been considered in recent literature: (in)fidelity, and sensitivity. We analyze optimal explanations with respect to both these measures, and while the optimal explanation for sensitivity is a vacuous constant explanation, the optimal explanation for infidelity is a novel combination of two popular explanation methods. By varying the perturbation distribution that defines infidelity, we obtain novel explanations by optimizing infidelity, which we show to out-perform existing explanations in both quantitative and qualitative measurements. Another salient question given these measures is how to modify any given explanation to have better values with respect to these measures. We propose a simple modification based on lowering sensitivity, and moreover show that when done appropriately, we could simultaneously improve both sensitivity as well as fidelity.
****

Manifold denoising by Nonlinear Robust Principal Component Analysis
He Lyu, Ningyu Sha, Shuyang Qin, Ming Yan, Yuying Xie, Rongrong Wang
This paper extends robust principal component analysis (RPCA) to nonlinear manifolds. Suppose that the observed data matrix is the sum of a sparse component and a component drawn from some low dimensional manifold. Is it possible to separate them by using similar ideas as RPCA? Is there any benefit in treating the manifold as a whole as opposed to treating each local region independently? We answer these two questions affirmatively by proposing and analyzing an optimization framework that separates the sparse component from the manifold under noisy data. Theoretical error bounds are provided when the tangent spaces of the manifold satisfy certain incoherence conditions. We also provide a near optimal choice of the tuning parameters for the proposed optimization formulation with the help of a new curvature estimation method. The efficacy of our method is demonstrated on both synthetic and real datasets.
****

Foundations of Comparison-Based Hierarchical Clustering
Debarghya Ghoshdastidar, Michaël Perrot, Ulrike von Luxburg
We address the classical problem of hierarchical clustering, but in a framework where one does not have access to a representation of the objects or their pairwise similarities. Instead, we assume that only a set of comparisons between objects is available, that is, statements of the form ``objects i and j are more similar than objects k and l.'' Such a scenario is commonly encountered in crowdsourcing applications. The focus of this work is to develop comparison-based hierarchical clustering algorithms that do not rely on the principles of ordinal embedding. We show that single and complete linkage are inherently comparison-based and we develop variants of average linkage. We provide statistical guarantees for the different methods under a planted hierarchical partition model. We also empirically demonstrate the performance of the proposed approaches on several datasets.
****

On the Accuracy of Influence Functions for Measuring Group Effects
Pang Wei W. Koh, Kai-Siang Ang, Hubert Teo, Percy S. Liang
Influence functions estimate the effect of removing a training point on a model without the need to retrain. They are based on a first-order Taylor approximation that is guaranteed to be accurate for sufficiently small changes to the model, and so are commonly used to study the effect of individual points in large datasets. However, we often want to study the effects of large groups of training points, e.g., to diagnose batch effects or apportion credit between different data sources. Removing such large groups can result in significant changes to the model. Are influence functions still accurate in this setting? In this paper, we find that across many different types of groups and for a range of real-world datasets, the predicted effect (using influence functions) of a group correlates surprisingly well with its actual effect, even if the absolute and relative errors are large. Our theoretical analysis shows that such strong correlation arises only under certain settings and need not hold in general, indicating that real-world datasets have particular properties that allow the influence approximation t

o be accurate.
****
Neural Similarity Learning
Weiyang Liu, Zhen Liu, James M. Rehg, Le Song
Inner product-based convolution has been the founding stone of convolutional neural networks (CNNs), enabling end-to-end learning of visual representation. By generalizing inner product with a bilinear matrix, we propose the neural similarity which serves as a learnable parametric similarity measure for CNNs. Neural similarity naturally generalizes the convolution and enhances flexibility. Further, we consider the neural similarity learning (NSL) in order to learn the neural similarity adaptively from training data. Specifically, we propose two different ways of learning the neural similarity: static NSL and dynamic NSL. Interestingly, dynamic neural similarity makes the CNN become a dynamic inference network. By regularizing the bilinear matrix, NSL can be viewed as learning the shape of kernel and the similarity measure simultaneously. We further justify the effectiveness of NSL with a theoretical viewpoint. Most importantly, NSL shows promising performance in visual recognition and few-shot learning, validating the superiority of NSL over the inner product-based convolution counterparts.
****
Multi-objective Bayesian optimisation with preferences over objectives
Majid Abdolshah, Alistair Shilton, Santu Rana, Sunil Gupta, Svetha Venkatesh
We present a multi-objective Bayesian optimisation algorithm that allows the user to express preference-order constraints on the objectives of the type objective A is more important than objective B. These preferences are defined based on the stability of the obtained solutions with respect to preferred objective functions. Rather than attempting to find a representative subset of the complete Pareto front, our algorithm selects those Pareto-optimal points that satisfy these constraints. We formulate a new acquisition function based on expected improvement in dominated hypervolume (EHI) to ensure that the subset of Pareto front satisfying the constraints is thoroughly explored. The hypervolume calculation is weighted by the probability of a point satisfying the constraints from a gradient Gaussian Process model. We demonstrate our algorithm on both synthetic and real-world problems.
****
Global Convergence of Least Squares EM for Demixing Two Log-Concave Densities
Wei Qian, Yuqian Zhang, Yudong Chen
This work studies the location estimation problem for a mixture of two rotation invariant log-concave densities. We demonstrate that Least Squares EM, a variant of the EM algorithm, converges to the true location parameter from a randomly initialized point. Moreover, we establish the explicit convergence rates and sample complexity bounds, revealing their dependence on the signal-to-noise ratio and the tail property of the log-concave distributions. Our analysis generalizes previous techniques for proving the convergence results of Gaussian mixtures, and highlights that an angle-decreasing property is sufficient for establishing global convergence for Least Squares EM.
****
The Case for Evaluating Causal Models Using Interventional Measures and Empirical Data
Amanda Gentzel, Dan Garant, David Jensen
Causal inference is central to many areas of artificial intelligence, including complex reasoning, planning, knowledge-base construction, robotics, explanation, and fairness. An active community of researchers develops and enhances algorithms that learn causal models from data, and this work has produced a series of impressive technical advances. However, evaluation techniques for causal modeling algorithms have remained somewhat primitive, limiting what we can learn from experimental studies of algorithm performance, constraining the types of algorithms and model representations that researchers consider, and creating a gap between theory and practice. We argue for more frequent use of evaluation techniques that examine interventional measures rather than structural or observational measures, and that evaluate those measures on empirical data rather than synthetic

data. We survey the current practice in evaluation and show that the techniques we recommend are rarely used in practice. We show that such techniques are feasible and that data sets are available to conduct such evaluations. We also show that these techniques produce substantially different results than using structural measures and synthetic data.
****
Spatially Aggregated Gaussian Processes with Multivariate Areal Outputs
Yusuke Tanaka, Toshiyuki Tanaka, Tomoharu Iwata, Takeshi Kurashima, Maya Okawa, Yasunori Akagi, Hiroyuki Toda
We propose a probabilistic model for inferring the multivariate function from multiple areal data sets with various granularities. Here, the areal data are observed not at location points but at regions. Existing regression-based models can only utilize the sufficiently fine-grained auxiliary data sets on the same domain (e.g., a city). With the proposed model, the functions for respective areal data sets are assumed to be a multivariate dependent Gaussian process (GP) that is modeled as a linear mixing of independent latent GPs. Sharing of latent GPs across multiple areal data sets allows us to effectively estimate the spatial correlation for each areal data set; moreover it can easily be extended to transfer learning across multiple domains. To handle the multivariate areal data, we design an observation model with a spatial aggregation process for each areal data set, which is an integral of the mixed GP over the corresponding region. By deriving the posterior GP, we can predict the data value at any location point by considering the spatial correlations and the dependences between areal data sets, simultaneously. Our experiments on real-world data sets demonstrate that our model can 1) accurately refine coarse-grained areal data, and 2) offer performance improvements by using the areal data sets from multiple domains.
****
First Exit Time Analysis of Stochastic Gradient Descent Under Heavy-Tailed Gradient Noise
Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, Gaël RICHARD
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.
****
Acceleration via Symplectic Discretization of High-Resolution Differential Equations
Bin Shi, Simon S. Du, Weijie Su, Michael I. Jordan
We study first-order optimization algorithms obtained by discretizing ordinary differential equations (ODEs) corresponding to Nesterov’s accelerated gradient methods (NAGs) and Polyak’s heavy-ball method. We consider three discretization schemes: symplectic Euler (S), explicit Euler (E) and implicit Euler (I) schemes. We show that the optimization algorithm generated by applying the symplectic scheme to a high-resolution ODE proposed by Shi et al. [2018] achieves the accelerated rate for minimizing both strongly convex function and convex function. On the other hand, the resulting algorithm either fails to achieve acceleration or is impractical when the scheme is implicit, the ODE is low-resolution, or the scheme is explicit.
****
Seeing the Wind: Visual Wind Speed Prediction with a Coupled Convolutional and Recurrent Neural Network
Jennifer Cardona, Michael Howland, John Dabiri
Wind energy resource quantification, air pollution monitoring, and weather forecasting all rely on rapid, accurate measurement of local wind conditions. Visual observations of the effects of wind---the swaying of trees and flapping of flags, for example---encode information regarding local wind conditions that can potentially be leveraged for visual anemometry that is inexpensive and ubiquitous. Here, we demonstrate a coupled convolutional neural network and recurrent neural network architecture that extracts the wind speed encoded in visually recorded flow-structure interactions of a flag and tree in naturally occurring wind. Predi

ctions for wind speeds ranging from 0.75-11 m/s showed agreement with measuremen ts from a cup anemometer on site, with a root-mean-squared error approaching the natural wind speed variability due to atmospheric turbulence. Generalizability of the network was demonstrated by successful prediction of wind speed based on recordings of other flags in the field and in a controlled wind tunnel test. Fur thermore, physics-based scaling of the flapping dynamics accurately predicts the dependence of the network performance on the video frame rate and duration.
****

Hyper-Graph-Network Decoders for Block Codes
Eliya Nachmani, Lior Wolf
Neural decoders were shown to outperform classical message passing techniques fo r short BCH codes. In this work, we extend these results to much larger families of algebraic block codes, by performing message passing with graph neural netw orks. The parameters of the sub-network at each variable-node in the Tanner grap h are obtained from a hypernetwork that receives the absolute values of the curr ent message as input. To add stability, we employ a simplified version of the ar ctanh activation that is based on a high order Taylor approximation of this acti vation function. Our results show that for a large number of algebraic block cod es, from diverse families of codes (BCH, LDPC, Polar), the decoding obtained wit h our method outperforms the vanilla belief propagation method as well as other learning techniques from the literature.
****

Sliced Gromov-Wasserstein
Vayer Titouan, Rémi Flamary, Nicolas Courty, Romain Tavenard, Laetitia Chapel
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.
****

Sparse Logistic Regression Learns All Discrete Pairwise Graphical Models
Shanshan Wu, Sujay Sanghavi, Alexandros G. Dimakis
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.
****

Coordinated hippocampal-entorhinal replay as structural inference
Talfan Evans, Neil Burgess
Constructing and maintaining useful representations of sensory experience is ess ential for reasoning about ones environment. High-level associative (topological) maps can be useful for efficient planning and are easily constructed from expe rience. Conversely, embedding new experiences within a metric structure allows t hem to be integrated with existing ones and novel associations to be implicitly inferred. Neurobiologically, the synaptic associations between hippocampal place cells and entorhinal grid cells are thought to represent associative and metric structures, respectively. Learning the place-grid cell associations can therefo re be interpreted as learning a mapping between these two spaces. Here, we show how this map could be constructed by probabilistic message-passing through the h ippocampal-entorhinal system, where messages are scheduled to reduce the propaga tion of redundant information. We propose that this offline inference correspond s to coordinated hippocampal-entorhinal replay during sharp wave ripples. Our re sults also suggest that the metric map will contain local distortions that refle ct the inferred structure of the environment according to associative experience , explaining observed grid deformations.
****

A Linearly Convergent Method for Non-Smooth Non-Convex Optimization on the Grass mannian with Applications to Robust Subspace and Dictionary Learning
Zhihui Zhu, Tianyu Ding, Daniel Robinson, Manolis Tsakiris, René Vidal
Minimizing a non-smooth function over the Grassmannian appears in many applicat ions in machine learning. In this paper we show that if the objective satisfies

a certain Riemannian regularity condition with respect to some point in the Grassmannian, then a Riemannian subgradient method with appropriate initialization and geometrically diminishing step size converges at a linear rate to that point. We show that for both the robust subspace learning method Dual Principal Component Pursuit (DPCP) and the Orthogonal Dictionary Learning (ODL) problem, the Riemannian regularity condition is satisfied with respect to appropriate points of interest, namely the subspace orthogonal to the sought subspace for DPCP and the orthonormal dictionary atoms for ODL. Consequently, we obtain in a unified framework significant improvements for the convergence theory of both methods.
****
Finite-time Analysis of Approximate Policy Iteration for the Linear Quadratic Regulator
Karl Krauth, Stephen Tu, Benjamin Recht
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 Impact of Regularization on High-dimensional Logistic Regression
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.
****
Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition
Jinwoo Choi, Chen Gao, Joseph C. E. Messou, Jia-Bin Huang
Human activities often occur in specific scene contexts, e.g., playing basketball on a basketball court. Training a model using existing video datasets thus inevitably captures and leverages such bias (instead of using the actual discriminative cues). The learned representation may not generalize well to new action classes or different tasks. In this paper, we propose to mitigate scene bias for video representation learning. Specifically, we augment the standard cross-entropy loss for action classification with 1) an adversarial loss for scene types and 2) a human mask confusion loss for videos where the human actors are masked out. These two losses encourage learning representations that are unable to predict the scene types and the correct actions when there is no evidence. We validate the effectiveness of our method by transferring our pre-trained model to three different tasks, including action classification, temporal localization, and spatio-temporal action detection. Our results show consistent improvement over the baseline model without debiasing.
****
(Nearly) Efficient Algorithms for the Graph Matching Problem on Correlated Random Graphs
Boaz Barak, Chi-Ning Chou, Zhixian Lei, Tselil Schramm, Yueqi Sheng
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.
****
Cross-sectional Learning of Extremal Dependence among Financial Assets
Xing Yan, Qi Wu, Wen 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.
****
Invert to Learn to Invert
Patrick Putzky, Max Welling

Iterative learning to infer approaches have become popular solvers for inverse problems. However, their memory requirements during training grow linearly with model depth, limiting in practice model expressiveness. In this work, we propose an iterative inverse model with constant memory that relies on invertible networks to avoid storing intermediate activations. As a result, the proposed approach allows us to train models with 400 layers on 3D volumes in an MRI image reconstruction task. In experiments on a public data set, we demonstrate that these deeper, and thus more expressive, networks perform state-of-the-art image reconstruction.
****

Metamers of neural networks reveal divergence from human perceptual systems
Jenelle Feather, Alex Durango, Ray Gonzalez, Josh McDermott
Deep neural networks have been embraced as models of sensory systems, instantiating representational transformations that appear to resemble those in the visual and auditory systems. To more thoroughly investigate their similarity to biological systems, we synthesized model metamers – stimuli that produce the same responses at some stage of a network’s representation. We generated model metamers for natural stimuli by performing gradient descent on a noise signal, matching the responses of individual layers of image and audio networks to a natural image or speech signal. The resulting signals reflect the invariances instantiated in the network up to the matched layer. We then measured whether model metamers were recognizable to human observers – a necessary condition for the model representations to replicate those of humans. Although model metamers from early network layers were recognizable to humans, those from deeper layers were not. Auditory model metamers became more human-recognizable with architectural modifications that reduced aliasing from pooling operations, but those from the deepest layers remained unrecognizable. We also used the metamer test to compare model representations. Cross-model metamer recognition dropped off for deeper layers, roughly at the same point that human recognition deteriorated, indicating divergence across model representations. The results reveal discrepancies between model and human representations, but also show how metamers can help guide model refinement and elucidate model representations.
****

Optimal Sparse Decision Trees
Xiyang Hu, Cynthia Rudin, Margo Seltzer
Decision tree algorithms have been among the most popular algorithms for interpretable (transparent) machine learning since the early 1980's. The problem that has plagued decision tree algorithms since their inception is their lack of optimality, or lack of guarantees of closeness to optimality: decision tree algorithms are often greedy or myopic, and sometimes produce unquestionably suboptimal models. Hardness of decision tree optimization is both a theoretical and practical obstacle, and even careful mathematical programming approaches have not been able to solve these problems efficiently. This work introduces the first practical algorithm for optimal decision trees for binary variables. The algorithm is a co-design of analytical bounds that reduce the search space and modern systems techniques, including data structures and a custom bit-vector library. We highlight possible steps to improving the scalability and speed of future generations of this algorithm based on insights from our theory and experiments.
****

Distinguishing Distributions When Samples Are Strategically Transformed
Hanrui Zhang, Yu Cheng, Vincent Conitzer
Often, a principal must make a decision based on data provided by an agent. Moreover, typically, that agent has an interest in the decision that is not perfectly aligned with that of the principal. Thus, the agent may have an incentive to select from or modify the samples he obtains before sending them to the principal. In other settings, the principal may not even be able to observe samples directly; instead, she must rely on signals that the agent is able to send based on the samples that he obtains, and he will choose these signals strategically.
****

Positive-Unlabeled Compression on the Cloud

Yixing Xu, Yunhe Wang, Hanting Chen, Kai Han, Chunjing XU, Dacheng Tao, Chang Xu
Many attempts have been done to extend the great success of convolutional neural networks (CNNs) achieved on high-end GPU servers to portable devices such as smart phones. Providing compression and acceleration service of deep learning models on the cloud is therefore of significance and is attractive for end users. However, existing network compression and acceleration approaches usually fine-tuning the svelte model by requesting the entire original training data (e.g. ImageNet), which could be more cumbersome than the network itself and cannot be easily uploaded to the cloud. In this paper, we present a novel positive-unlabeled (PU) setting for addressing this problem. In practice, only a small portion of the original training set is required as positive examples and more useful training examples can be obtained from the massive unlabeled data on the cloud through a PU classifier with an attention based multi-scale feature extractor. We further introduce a robust knowledge distillation (RKD) scheme to deal with the class imbalance problem of these newly augmented training examples. The superiority of the proposed method is verified through experiments conducted on the benchmark models and datasets. We can use only 8% of uniformly selected data from the ImageNet to obtain an efficient model with comparable performance to the baseline ResNet-34.
****
Nonparametric Contextual Bandits in Metric Spaces with Unknown Metric
Nirandika Wanigasekara, Christina 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.
****
Staying up to Date with Online Content Changes Using Reinforcement Learning for Scheduling
Andrey Kolobov, Yuval Peres, Cheng Lu, Eric J. Horvitz
From traditional Web search engines to virtual assistants and Web accelerators, services that rely on online information need to continually keep track of remote content changes by explicitly requesting content updates from remote sources (e.g., web pages). We propose a novel optimization objective for this setting that has several practically desirable properties, and efficient algorithms for it with optimality guarantees even in the face of mixed content change observability and initially unknown change model parameters. Experiments on 18.5M URLs crawled daily for 14 weeks show significant advantages of this approach over prior art.
****
Interlaced Greedy Algorithm for Maximization of Submodular Functions in Nearly Linear Time
Alan Kuhnle
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 unified variance-reduced accelerated gradient method for convex optimization
Guanghui Lan, Zhize Li, Yi Zhou
We propose a novel randomized incremental gradient algorithm, namely, VAriance-Reduced Accelerated Gradient (Varag), for finite-sum optimization. Equipped with a unified step-size policy that adjusts itself to the value of the conditional number, Varag exhibits the unified optimal rates of convergence for solving smooth convex finite-sum problems directly regardless of their strong convexity. Moreover, Varag is the first accelerated randomized incremental gradient method that benefits from the strong convexity of the data-fidelity term to achieve the optimal linear convergence. It also establishes an optimal linear rate of convergence for solving a wide class of problems only satisfying a certain error bound condition rather than strong convexity. Varag can also be extended to solve stocha

stic finite-sum problems.
****
SSRGD: Simple Stochastic Recursive Gradient Descent for Escaping Saddle Points
Zhize 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-auth
ors prior to requesting a name change in the electronic proceedings.
****
This Looks Like That: Deep Learning for Interpretable Image Recognition
Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, Jonathan K. Su
When we are faced with challenging image classification tasks, we often explain
our reasoning by dissecting the image, and pointing out prototypical aspects of
one class or another. The mounting evidence for each of the classes helps us mak
e our final decision. In this work, we introduce a deep network architecture --
prototypical part network (ProtoPNet), that reasons in a similar way: the networ
k dissects the image by finding prototypical parts, and combines evidence from t
he prototypes to make a final classification. The model thus reasons in a way th
at is qualitatively similar to the way ornithologists, physicians, and others wo
uld explain to people on how to solve challenging image classification tasks. Th
e network uses only image-level labels for training without any annotations for
parts of images. We demonstrate our method on the CUB-200-2011 dataset and the S
tanford Cars dataset. Our experiments show that ProtoPNet can achieve comparable
 accuracy with its analogous non-interpretable counterpart, and when several Pro
toPNets are combined into a larger network, it can achieve an accuracy that is o
n par with some of the best-performing deep models. Moreover, ProtoPNet provides
 a level of interpretability that is absent in other interpretable deep models.
****
Online EXP3 Learning in Adversarial Bandits with Delayed Feedback
Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, Jose Blanchet
Consider a player that in each of T rounds chooses one of K arms. An adversary c
hooses the cost of each arm in a bounded interval, and a sequence of feedback de
lays \left{ d{t}\right} that are unknown to the player. After picking arm a{t}
at round t, the player receives the cost of playing this arm d{t} rounds later.
In cases where t+d{t}>T, this feedback is simply missing. We prove that the EXP3
 algorithm (that uses the delayed feedback upon its arrival) achieves a regret o
f O\left(\sqrt{\ln K\left(KT+\sum{t=1}^{T}d{t}\right)}\right). For the case wher
e \sum{t=1}^{T}d{t} and T are unknown, we propose a novel doubling trick for onl
ine learning with delays and prove that this adaptive EXP3 achieves a regret of
O\left(\sqrt{\ln K\left(K^{2}T+\sum{t=1}^{T}d{t}\right)}\right). We then conside
r a two player zero-sum game where players experience asynchronous delays. We sh
ow that even when the delays are large enough such that players no longer enjoy
the “no-regret property”, (e.g., where d{t}=O\left(t\log t\right)) the ergodic a
verage of the strategy profile still converges to the set of Nash equilibria of
the game. The result is made possible by choosing an adaptive step size \eta{t}
that is not summable but is square summable, and proving a “weighted regret boun
d” for this general case.
****
Phase Transitions and Cyclic Phenomena in Bandits with Switching Constraints
David Simchi-Levi, Yunzong Xu
We consider the classical stochastic multi-armed bandit problem with a constrain
t on the total cost incurred by switching between actions. Under the unit switch
ing cost structure, where the constraint limits the total number of switches, we
 prove matching upper and lower bounds on regret and provide near-optimal algori
thms for this problem. Surprisingly, we discover phase transitions and cyclic ph
enomena of the optimal regret. That is, we show that associated with the multi-a
rmed bandit problem, there are equal-length phases defined by the number of arms
 and switching costs, where the regret upper and lower bounds in each phase rema
in the same and drop significantly between phases. The results enable us to full
y characterize the trade-off between regret and incurred switching cost in the s

tochastic multi-armed bandit problem, contributing new insights to this fundamental problem. Under the general switching cost structure, our analysis reveals a surprising connection between the bandit problem and the shortest Hamiltonian path problem.
****
Learning Dynamics of Attention: Human Prior for Interpretable Machine Reasoning
Wonjae Kim, Yoonho Lee
Without relevant human priors, neural networks may learn uninterpretable features. We propose Dynamics of Attention for Focus Transition (DAFT) as a human prior for machine reasoning. DAFT is a novel method that regularizes attention-based reasoning by modelling it as a continuous dynamical system using neural ordinary differential equations. As a proof of concept, we augment a state-of-the-art visual reasoning model with DAFT. Our experiments reveal that applying DAFT yields similar performance to the original model while using fewer reasoning steps, showing that it implicitly learns to skip unnecessary steps. We also propose a new metric, Total Length of Transition (TLT), which represents the effective reasoning step size by quantifying how much a given model's focus drifts while reasoning about a question. We show that adding DAFT results in lower TLT, demonstrating that our method indeed obeys the human prior towards shorter reasoning paths in addition to producing more interpretable attention maps.
****
Provable Certificates for Adversarial Examples: Fitting a Ball in the Union of Polytopes
Matt Jordan, Justin Lewis, Alexandros G. Dimakis
We propose a novel method for computing exact pointwise robustness of deep neural networks for all convex lp norms. Our algorithm, GeoCert, finds the largest lp ball centered at an input point x0, within which the output class of a given neural network with ReLU nonlinearities remains unchanged. We relate the problem of computing pointwise robustness of these networks to that of computing the maximum norm ball with a fixed center that can be contained in a non-convex polytope. This is a challenging problem in general, however we show that there exists an efficient algorithm to compute this for polyhedral complices. Further we show that piecewise linear neural networks partition the input space into a polyhedral complex. Our algorithm has the ability to almost immediately output a nontrivial lower bound to the pointwise robustness which is iteratively improved until it ultimately becomes tight. We empirically show that our approach generates a distance lower bounds that are tighter compared to prior work, under moderate time constraints.
****
Fast Parallel Algorithms for Statistical Subset Selection Problems
Sharon Qian, Yaron Singer
In this paper, we propose a new framework for designing fast parallel algorithms for fundamental statistical subset selection tasks that include feature selection and experimental design. Such tasks are known to be weakly submodular and are amenable to optimization via the standard greedy algorithm. Despite its desirable approximation guarantees, however, the greedy algorithm is inherently sequential and in the worst case, its parallel runtime is linear in the size of the data.
Recently, there has been a surge of interest in a parallel optimization technique called adaptive sampling which produces solutions with desirable approximation guarantees for submodular maximization in exponentially faster parallel runtime. Unfortunately, we show that for general weakly submodular functions such accelerations are impossible. The major contribution in this paper is a novel relaxation of submodularity which we call differential submodularity. We first prove that differential submodularity characterizes objectives like feature selection and experimental design. We then design an adaptive sampling algorithm for differentially submodular functions whose parallel runtime is logarithmic in the si

ze of the data and achieves strong approximation guarantees. Through experiments, we show the algorithm's performance is competitive with state-of-the-art methods and obtains dramatic speedups for feature selection and experimental design problems.
****

On Lazy Training in Differentiable Programming
Lénaïc Chizat, Edouard Oyallon, Francis Bach
In a series of recent theoretical works, it was shown that strongly over-parameterized neural networks trained with gradient-based methods could converge exponentially fast to zero training loss, with their parameters hardly varying. In this work, we show that this lazy training'' phenomenon is not specific to over-parameterized neural networks, and is due to a choice of scaling, often implicit, that makes the model behave as its linearization around the initialization, thus yielding a model equivalent to learning with positive-definite kernels. Through a theoretical analysis, we exhibit various situations where this phenomenon arises in non-convex optimization and we provide bounds on the distance between the lazy and linearized optimization paths. Our numerical experiments bring a critical note, as we observe that the performance of commonly used non-linear deep convolutional neural networks in computer vision degrades when trained in the lazy regime. This makes it unlikely thatlazy training'' is behind the many successes of neural networks in difficult high dimensional tasks.
****

Estimating Convergence of Markov chains with L-Lag Couplings
Niloy Biswas, Pierre E. Jacob, Paul Vanetti
Markov chain Monte Carlo (MCMC) methods generate samples that are asymptotically distributed from a target distribution of interest as the number of iterations goes to infinity. Various theoretical results provide upper bounds on the distance between the target and marginal distribution after a fixed number of iterations. These upper bounds are on a case by case basis and typically involve intractable quantities, which limits their use for practitioners. We introduce L-lag couplings to generate computable, non-asymptotic upper bound estimates for the total variation or the Wasserstein distance of general Markov chains. We apply L-lag couplings to the tasks of (i) determining MCMC burn-in, (ii) comparing different MCMC algorithms with the same target, and (iii) comparing exact and approximate MCMC. Lastly, we (iv) assess the bias of sequential Monte Carlo and self-normalized importance samplers.
****

Efficient Regret Minimization Algorithm for Extensive-Form Correlated Equilibrium
Gabriele Farina, Chun Kai Ling, Fei Fang, Tuomas Sandholm
Self-play methods based on regret minimization have become the state of the art for computing Nash equilibria in large two-players zero-sum extensive-form games. These methods fundamentally rely on the hierarchical structure of the players' sequential strategy spaces to construct a regret minimizer that recursively minimizes regret at each decision point in the game tree. In this paper, we introduce the first efficient regret minimization algorithm for computing extensive-form correlated equilibria in large two-player general-sum games with no chance moves. Designing such an algorithm is significantly more challenging than designing one for the Nash equilibrium counterpart, as the constraints that define the space of correlation plans lack the hierarchical structure and might even form cycles. We show that some of the constraints are redundant and can be excluded from consideration, and present an efficient algorithm that generates the space of extensive-form correlation plans incrementally from the remaining constraints. This structural decomposition is achieved via a special convexity-preserving operation that we coin scaled extension. We show that a regret minimizer can be designed for a scaled extension of any two convex sets, and that from the decomposition we then obtain a global regret minimizer. Our algorithm produces feasible iterates. Experiments show that it significantly outperforms prior approaches and for larger problems it is the only viable option.
****

Using Embeddings to Correct for Unobserved Confounding in Networks
Victor Veitch, Yixin Wang, David Blei
We consider causal inference in the presence of unobserved confounding. We study the case where a proxy is available for the unobserved confounding in the form of a network connecting the units. For example, the link structure of a social network carries information about its members. We show how to effectively use the proxy to do causal inference. The main idea is to reduce the causal estimation problem to a semi-supervised prediction of both the treatments and outcomes. Networks admit high-quality embedding models that can be used for this semi-supervised prediction. We show that the method yields valid inferences under suitable (weak) conditions on the quality of the predictive model. We validate the method with experiments on a semi-synthetic social network dataset.
****
Towards Practical Alternating Least-Squares for CCA
Zhiqiang Xu, Ping Li
Alternating least-squares (ALS) is a simple yet effective solver for canonical correlation analysis (CCA). In terms of ease of use, ALS is arguably practitioners' first choice. Despite recent provably guaranteed variants, the empirical performance often remains unsatisfactory. To promote the practical use of ALS for CCA, we propose truly alternating least-squares. Instead of approximately solving two independent linear systems, in each iteration, it simply solves two coupled linear systems of half the size. It turns out that this coupling procedure is able to bring significant performance improvements in practice. Inspired by accelerated power method, we further propose faster alternating least-squares, where momentum terms are introduced into the update equations. Both algorithms enjoy linear convergence. To make faster ALS even more practical, we put forward adaptive alternating least-squares to avoid tuning the momentum parameter, which is as easy to use as the plain ALS while retaining advantages of the fast version. Experiments on several datasets empirically demonstrate the superiority of the proposed algorithms to recent variants.
****
Neural Multisensory Scene Inference
Jae Hyun Lim, Pedro O. O. Pinheiro, Negar Rostamzadeh, Chris Pal, Sungjin Ahn
For embodied agents to infer representations of the underlying 3D physical world they inhabit, they should efficiently combine multisensory cues from numerous trials, e.g., by looking at and touching objects. Despite its importance, multisensory 3D scene representation learning has received less attention compared to the unimodal setting. In this paper, we propose the Generative Multisensory Network (GMN) for learning latent representations of 3D scenes which are partially observable through multiple sensory modalities. We also introduce a novel method, called the Amortized Product-of-Experts, to improve the computational efficiency and the robustness to unseen combinations of modalities at test time. Experimental results demonstrate that the proposed model can efficiently infer robust modality-invariant 3D-scene representations from arbitrary combinations of modalities and perform accurate cross-modal generation.
To perform this exploration we have also developed a novel multi-sensory simulation environment for embodied agents.
****
Emergence of Object Segmentation in Perturbed Generative Models
Adam Bielski, Paolo Favaro
We introduce a novel framework to build a model that can learn how to segment objects from a collection of images without any human annotation. Our method builds on the observation that the location of object segments can be perturbed locally relative to a given background without affecting the realism of a scene. Our approach is to first train a generative model of a layered scene. The layered representation consists of a background image, a foreground image and the mask of the foreground. A composite image is then obtained by overlaying the masked foreground image onto the background. The generative model is trained in an adversarial fashion against a discriminator, which forces the generative model to produce realistic composite images. To force the generator to learn a representation w

here the foreground layer corresponds to an object, we perturb the output of the generative model by introducing a random shift of both the foreground image and mask relative to the background. Because the generator is unaware of the shift before computing its output, it must produce layered representations that are realistic for any such random perturbation. Finally, we learn to segment an image by defining an autoencoder consisting of an encoder, which we train, and the pre-trained generator as the decoder, which we freeze. The encoder maps an image to a feature vector, which is fed as input to the generator to give a composite image matching the original input image. Because the generator outputs an explicit layered representation of the scene, the encoder learns to detect and segment objects. We demonstrate this framework on real images of several object categories.
****
Learning Transferable Graph Exploration
Hanjun Dai, Yujia Li, Chenglong Wang, Rishabh Singh, Po-Sen Huang, Pushmeet Kohli
This paper considers the problem of efficient exploration of unseen environments, a key challenge in AI. We propose a learning to explore' framework where we learn a policy from a distribution of environments. At test time, presented with an unseen environment from the same distribution, the policy aims to generalize the exploration strategy to visit the maximum number of unique states in a limited number of steps. We particularly focus on environments with graph-structured state-spaces that are encountered in many important real-world applications like software testing and map building.
We formulate this task as a reinforcement learning problem where theexploration' agent is rewarded for transitioning to previously unseen environment states and employ a graph-structured memory to encode the agent's past trajectory. Experimental results demonstrate that our approach is extremely effective for exploration of spatial maps; and when applied on the challenging problems of coverage-guided software-testing of domain-specific programs and real-world mobile applications, it outperforms methods that have been hand-engineered by human experts.
****
On the Optimality of Perturbations in Stochastic and Adversarial Multi-armed Bandit Problems
Baekjin Kim, Ambuj Tewari
We investigate the optimality of perturbation based algorithms in the stochastic and adversarial multi-armed bandit problems. For the stochastic case, we provide a unified regret analysis for both sub-Weibull and bounded perturbations when rewards are sub-Gaussian. Our bounds are instance optimal for sub-Weibull perturbations with parameter 2 that also have a matching lower tail bound, and all bounded support perturbations where there is sufficient probability mass at the extremes of the support. For the adversarial setting, we prove rigorous barriers against two natural solution approaches using tools from discrete choice theory and extreme value theory. Our results suggest that the optimal perturbation, if it exists, will be of Frechet-type.
****
Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions
Gabriele Farina, 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.
****
A Fourier Perspective on Model Robustness in Computer Vision
Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, Justin Gilmer
Achieving robustness to distributional shift is a longstanding and challenging goal of computer vision. Data augmentation is a commonly used approach for improving robustness, however robustness gains are typically not uniform across corruption types. Indeed increasing performance in the presence of random noise is oft

en met with reduced performance on other corruptions such as contrast change. Un derstanding when and why these sorts of trade-offs occur is a crucial step towar ds mitigating them. Towards this end, we investigate recently observed trade-off s caused by Gaussian data augmentation and adversarial training. We find that bo th methods improve robustness to corruptions that are concentrated in the high f requency domain while reducing robustness to corruptions that are concentrated i n the low frequency domain. This suggests that one way to mitigate these trade-o ffs via data augmentation is to use a more diverse set of augmentations.
Towards this end we observe that AutoAugment, a recently proposed data augmentat ion policy optimized for clean accuracy, achieves state-of-the-art robustness on the CIFAR-10-C benchmark.
****
Two Generator Game: Learning to Sample via Linear Goodness-of-Fit Test
Lizhong Ding, Mengyang Yu, Li Liu, Fan Zhu, Yong Liu, Yu Li, Ling Shao
Learning the probability distribution of high-dimensional data is a challenging problem. To solve this problem, we formulate a deep energy adversarial network (DEAN), which casts the energy model learned from real data into an optimization of a goodness-of-fit (GOF) test statistic. DEAN can be interpreted as a GOF game between two generative networks, where one explicit generative network learns a n energy-based distribution that fits the real data, and the other implicit gene rative network is trained by minimizing a GOF test statistic between the energy-based distribution and the generated data, such that the underlying distribution of the generated data is close to the energy-based distribution. We design a tw o-level alternative optimization procedure to train the explicit and implicit ge nerative networks, such that the hyper-parameters can also be automatically lear ned. Experimental results show that DEAN achieves high quality generations compa red to the state-of-the-art approaches.
****
Fixing Implicit Derivatives: Trust-Region Based Learning of Continuous Energy Fu nctions
Chris Russell, Matteo Toso, Neill Campbell
We present a new technique for the learning of continuous energy functions that we refer to as Wibergian Learning. One common approach to inverse problems is to cast them as an energy minimisation problem, where the minimum cost solution found is used as an estimator of hidden parameters. Our new approach formally characterises the dependency between weights that control the shape of the energy function, and the location of minima, by describing minima as fixed points of optimisation methods. This allows for the use of gradient-based end-to
-
end training to integrate deep-learning and the classical inverse problem method s.
We show how our approach can be applied to obtain state-of-the-art results in th e
diverse applications of tracker fusion and multiview 3D reconstruction.
****
Correlation Clustering with Adaptive Similarity Queries
Marco Bressan, Nicolò Cesa-Bianchi, Andrea Paudice, Fabio Vitale
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.
****
Deep imitation learning for molecular inverse problems
Eric Jonas
Many measurement modalities arise from well-understood physical processes and re sult in information-rich but difficult-to-interpret data. Much of this data stil l requires laborious human interpretation. This is the case in nuclear magnetic resonance (NMR) spectroscopy, where the observed spectrum of a molecule provides a distinguishing fingerprint of its bond structure. Here we solve the resulting inverse problem: given a molecular formula and a spectrum, can we infer the che

mical structure? We show for a wide variety of molecules we can quickly compute the correct molecular structure, and can detect with reasonable certainty when our method cannot. We treat this as a problem of graph-structured prediction, where armed with per-vertex information on a subset of the vertices, we infer the edges and edge types. We frame the problem as a Markov decision process (MDP) and incrementally construct molecules one bond at a time, training a deep neural network via imitation learning, where we learn to imitate a subisomorphic oracle which knows which remaining bonds are correct. Our method is fast, accurate, and is the first among recent chemical-graph generation approaches to exploit per-vertex information and generate graphs with vertex constraints. Our method points the way towards automation of molecular structure identification and potentially active learning for spectroscopy.
****
Ease-of-Teaching and Language Structure from Emergent Communication
Fushan Li, Michael Bowling
Artificial agents have been shown to learn to communicate when needed to complete a cooperative task. Some level of language structure (e.g., compositionality) has been found in the learned communication protocols. This observed structure is often the result of specific environmental pressures during training. By introducing new agents periodically to replace old ones, sequentially and within a population, we explore such a new pressure — ease of teaching — and show its impact on the structure of the resulting language.
****
Practical Differentially Private Top-k Selection with Pay-what-you-get Composition
David Durfee, Ryan M. Rogers
We study the problem of top-k selection over a large domain universe subject to user-level differential privacy. Typically, the exponential mechanism or report noisy max are the algorithms used to solve this problem. However, these algorithms require querying the database for the count of each domain element. We focus on the setting where the data domain is unknown, which is different than the setting of frequent itemsets where an apriori type algorithm can help prune the space of domain elements to query. We design algorithms that ensures (approximate) differential privacy and only needs access to the true top-k' elements from the data for any chosen k' ≥ k. This is a highly desirable feature for making differential privacy practical, since the algorithms require no knowledge of the domain. We consider both the setting where a user's data can modify an arbitrary number of counts by at most 1, i.e. unrestricted sensitivity, and the setting where a user's data can modify at most some small, fixed number of counts by at most 1, i.e. restricted sensitivity. Additionally, we provide a pay-what-you-get privacy composition bound for our algorithms. That is, our algorithms might return fewer than k elements when the top-k elements are queried, but the overall privacy budget only decreases by the size of the outcome set.
****
A Communication Efficient Stochastic Multi-Block Alternating Direction Method of Multipliers
Hao Yu
The alternating direction method of multipliers (ADMM) has recently received tremendous interests for distributed large scale optimization in machine learning, statistics, multi-agent networks and related applications. In this paper, we propose a new parallel multi-block stochastic ADMM for distributed stochastic optimization, where each node is only required to perform simple stochastic gradient descent updates. The proposed ADMM is fully parallel, can solve problems with arbitrary block structures, and has a convergence rate comparable to or better than existing state-of-the-art ADMM methods for stochastic optimization. Existing stochastic (or deterministic) ADMMs require each node to exchange its updated primal variables across nodes at each iteration and hence cause significant amount of communication overhead. Existing ADMMs require roughly the same number of inter-node communication rounds as the number of in-node computation rounds. In contrast, the number of communication rounds required by our new ADMM is only the s

quare root of the number of computation rounds.
****
Distributed estimation of the inverse Hessian by determinantal averaging
Michal Derezinski, Michael W. Mahoney
In distributed optimization and distributed numerical linear algebra,
we often encounter an inversion bias: if we want to compute a
quantity that depends on the inverse of a sum of distributed matrices,
then the sum of the inverses does not equal the inverse of the sum.
An example of this occurs in distributed Newton's method, where we
wish to compute (or implicitly work with) the inverse Hessian
multiplied by the gradient.
In this case, locally computed estimates are biased, and so taking a
uniform average will not recover the correct solution.
To address this, we propose determinantal averaging, a new
approach for correcting the inversion bias.
This approach involves reweighting the local estimates of the Newton's
step proportionally to the determinant of the local Hessian estimate,
and then averaging them together to obtain an improved global
estimate. This method provides the first known distributed Newton step that is
asymptotically consistent, i.e., it recovers the exact step in
the limit as the number of distributed partitions grows to infinity.
To show this, we develop new expectation identities and moment bounds
for the determinant and adjugate of a random matrix.
Determinantal averaging can be applied not only to Newton's method,
but to computing any quantity that is a linear tranformation of a
matrix inverse, e.g., taking a trace of the inverse covariance matrix,
which is used in data uncertainty quantification.
****
muSSP: Efficient Min-cost Flow Algorithm for Multi-object Tracking
Congchao Wang, Yizhi Wang, Yinxue Wang, Chiung-Ting Wu, Guoqiang Yu
Min-cost flow has been a widely used paradigm for solving data association probl
ems in multi-object tracking (MOT). However, most existing methods of solving mi
n-cost flow problems in MOT are either direct adoption or slight modifications o
f generic min-cost flow algorithms, yielding sub-optimal computation efficiency
and holding the applications back from larger scale of problems. In this paper,
by exploiting the special structures and properties of the graphs formulated in
MOT problems, we develop an efficient min-cost flow algorithm, namely, minimum-u
pdate Successive Shortest Path (muSSP). muSSP is proved to provide exact optimal
 solution and we demonstrated its efficiency through 40 experiments on five MOT
datasets with various object detection results and a number of graph designs. mu
SSP is always the most efficient in each experiment compared to the three peer s
olvers, improving the efficiency by 5 to 337 folds relative to the best competin
g algorithm and averagely 109 to 4089 folds to each of the three peer methods.
****
Invertible Convolutional Flow
Mahdi Karami, Dale Schuurmans, Jascha Sohl-Dickstein, Laurent Dinh, Daniel Duckw
orth
Normalizing flows can be used to construct high quality generative probabilistic
models, but training and sample generation require repeated evaluation of Jacobi
an determinants and function inverses. To make such computations feasible, curre
nt approaches employ highly constrained architectures that produce diagonal, tri
angular, or low rank Jacobian matrices. As an alternative, we investigate a set
 of novel normalizing flows based on the circular and symmetric convolutions. W
e show that these transforms admit efficient Jacobian determinant computation an
d inverse mapping (deconvolution) in O(N log N) time. Additionally, element-wise
 multiplication, widely used in normalizing flow architectures, can be combined
with these transforms to increase modeling flexibility.   We further propose an
analytic approach to designing nonlinear elementwise bijectors that induce speci
al properties in the intermediate layers, by implicitly introducing specific reg
ularizers in the loss. We show that these transforms allow more effective normal

izing flow models to be developed for generative image models.
****

Controlling Neural Level Sets
Matan Atzmon, Niv Haim, Lior Yariv, Ofer Israelov, Haggai Maron, Yaron Lipman
The level sets of neural networks represent fundamental properties such as decision boundaries of classifiers and are used to model non-linear manifold data such as curves and surfaces. Thus, methods for controlling the neural level sets could find many applications in machine learning.
****

Learning GANs and Ensembles Using Discrepancy
Ben Adlam, Corinna Cortes, Mehryar Mohri, Ningshan Zhang
Generative adversarial networks (GANs) generate data based on minimizing a divergence between two distributions. The choice of that divergence is therefore critical. We argue that the divergence must take into account the hypothesis set and the loss function used in a subsequent learning task, where the data generated by a GAN serves for training. Taking that structural information into account is also important to derive generalization guarantees. Thus, we propose to use the discrepancy measure, which was originally introduced for the closely related problem of domain adaptation and which precisely takes into account the hypothesis set and the loss function. We show that discrepancy admits favorable properties for training GANs and prove explicit generalization guarantees. We present efficient algorithms using discrepancy for two tasks: training a GAN directly, namely DGAN, and mixing previously trained generative models, namely EDGAN. Our experiments on toy examples and several benchmark datasets show that DGAN is competitive with other GANs and that EDGAN outperforms existing GAN ensembles, such as AdaGAN.
****

Neural Relational Inference with Fast Modular Meta-learning
Ferran Alet, Erica Weng, Tomás Lozano-Pérez, Leslie Pack Kaelbling
Graph neural networks (GNNs) are effective models for many dynamical systems consisting of entities and relations. Although most GNN applications assume a single type of entity and relation, many situations involve multiple types of interactions. Relational inference is the problem of inferring these interactions and learning the dynamics from observational data. We frame relational inference as a modular meta-learning problem, where neural modules are trained to be composed in different ways to solve many tasks. This meta-learning framework allows us to implicitly encode time invariance and infer relations in context of one another rather than independently, which increases inference capacity. Framing inference as the inner-loop optimization of meta-learning leads to a model-based approach that is more data-efficient and capable of estimating the state of entities that we do not observe directly, but whose existence can be inferred from their effect on observed entities. To address the large search space of graph neural network compositions, we meta-learn a proposal function that speeds up the inner-loop simulated annealing search within the modular meta-learning algorithm, providing two orders of magnitude increase in the size of problems that can be addressed.
****

Identification of Conditional Causal Effects under Markov Equivalence
Amin Jaber, Jiji Zhang, Elias Bareinboim
Causal identification is the problem of deciding whether a post-interventional distribution is computable from a combination of qualitative knowledge about the data-generating process, which is encoded in a causal diagram, and an observational distribution. A generalization of this problem restricts the qualitative knowledge to a class of Markov equivalent causal diagrams, which, unlike a single, fully-specified causal diagram, can be inferred from the observational distribution.
Recent work by (Jaber et al., 2019a) devised a complete algorithm for the identification of unconditional causal effects given a Markov equivalence class of causal diagrams. However, there are identifiable conditional causal effects that cannot be handled by that algorithm. In this work, we derive an algorithm to ident

ify conditional effects, which are particularly useful for evaluating conditiona l plans or policies.
****
Learning to Predict Layout-to-image Conditional Convolutions for Semantic Image Synthesis
Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, hongsheng Li
Semantic image synthesis aims at generating photorealistic images from semantic layouts. Previous approaches with conditional generative adversarial networks (G AN) show state-of-the-art performance on this task, which either feed the semant ic label maps as inputs to the generator, or use them to modulate the activation s in normalization layers via affine transformations. We argue that convolutiona l kernels in the generator should be aware of the distinct semantic labels at di fferent locations when generating images. In order to better exploit the semanti c layout for the image generator, we propose to predict convolutional kernels co nditioned on the semantic label map to generate the intermediate feature maps fr om the noise maps and eventually generate the images. Moreover, we propose a fea ture pyramid semantics-embedding discriminator, which is more effective in enhan cing fine details and semantic alignments between the generated images and the i nput semantic layouts than previous multi-scale discriminators. We achieve state -of-the-art results on both quantitative metrics and subjective evaluation on va rious semantic segmentation datasets, demonstrating the effectiveness of our app roach.
****
Average Case Column Subset Selection for Entrywise ℓ_1-Norm Loss
Zhao Song, David Woodruff, Peilin Zhong
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.
****
Piecewise Strong Convexity of Neural Networks
Tristan Milne
We study the loss surface of a feed-forward neural network with ReLU non-lineari ties, regularized with weight decay. We show that the regularized loss function is piecewise strongly convex on an important open set which contains, under some conditions, all of its global minimizers. This is used to prove that local mini ma of the regularized loss function in this set are isolated, and that every dif ferentiable critical point in this set is a local minimum, partially addressing an open problem given at the Conference on Learning Theory (COLT) 2015; our resu lt is also applied to linear neural networks to show that with weight decay regu larization, there are no non-zero critical points in a norm ball obtaining train ing error below a given threshold. We also include an experimental section where we validate our theoretical work and show that the regularized loss function is almost always piecewise strongly convex when restricted to stochastic gradient descent trajectories for three standard image classification problems.
****
No Pressure! Addressing the Problem of Local Minima in Manifold Learning Algorit hms
Max Vladymyrov
Nonlinear embedding manifold learning methods provide invaluable visual insights into a structure of high-dimensional data. However, due to a complicated noncon vex objective function, these methods can easily get stuck in local minima and t heir embedding quality can be poor. We propose a natural extension to several ma nifold learning methods aimed at identifying pressured points, i.e. points stuck in the poor local minima and have poor embedding quality. We show that the obje ctive function can be decreased by temporarily allowing these points to make use of an extra dimension in the embedding space. Our method is able to improve the objective function value of existing methods even after they get stuck in a poo r local minimum.
****

Approximate Inference Turns Deep Networks into Gaussian Processes
Mohammad Emtiyaz E. Khan, Alexander Immer, Ehsan Abedi, Maciej Korzepa
Deep neural networks (DNN) and Gaussian processes (GP) are two powerful models with several theoretical connections relating them, but the relationship between their training methods is not well understood. In this paper, we show that certain Gaussian posterior approximations for Bayesian DNNs are equivalent to GP posteriors. This enables us to relate solutions and iterations of a deep-learning algorithm to GP inference. As a result, we can obtain a GP kernel and a nonlinear feature map while training a DNN. Surprisingly, the resulting kernel is the neural tangent kernel. We show kernels obtained on real datasets and demonstrate the use of the GP marginal likelihood to tune hyperparameters of DNNs. Our work aims to facilitate further research on combining DNNs and GPs in practical settings.
****
Elliptical Perturbations for Differential Privacy
Matthew Reimherr, Jordan Awan
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.
****
Inherent Tradeoffs in Learning Fair Representations
Han Zhao, Geoff Gordon
With the prevalence of machine learning in high-stakes applications, especially the ones regulated by anti-discrimination laws or societal norms, it is crucial to ensure that the predictive models do not propagate any existing bias or discrimination. Due to the ability of deep neural nets to learn rich representations, recent advances in algorithmic fairness have focused on learning fair representations with adversarial techniques to reduce bias in data while preserving utility simultaneously. In this paper, through the lens of information theory, we provide the first result that quantitatively characterizes the tradeoff between demographic parity and the joint utility across different population groups. Specifically, when the base rates differ between groups, we show that any method aiming to learn fair representations admits an information-theoretic lower bound on the joint error across these groups. To complement our negative results, we also prove that if the optimal decision functions across different groups are close, then learning fair representations leads to an alternative notion of fairness, known as the accuracy parity, which states that the error rates are close between groups. Finally, our theoretical findings are also confirmed empirically on real-world datasets.
****
SGD on Neural Networks Learns Functions of Increasing Complexity
Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, Haofeng Zhang
We perform an experimental study of the dynamics of Stochastic Gradient Descent (SGD) in learning deep neural networks for several real and synthetic classification tasks.
We show that in the initial epochs, almost all of the performance improvement of the classifier obtained by SGD can be explained by a linear classifier.
More generally, we give evidence for the hypothesis that, as iterations progress, SGD learns functions of increasing complexity. This hypothesis can be helpful in explaining why SGD-learned classifiers tend to generalize well even in the over-parameterized regime.
We also show that the linear classifier learned in the initial stages is ``retained'' throughout the execution even if training is continued to the point of zero training error, and complement this with a theoretical result in a simplified model.
Key to our work is a new measure of
how well one classifier explains the performance of another, based on conditional mutual information.

****
Online Continuous Submodular Maximization: From Full-Information to Bandit Feedb
ack
Mingrui Zhang, Lin Chen, Hamed Hassani, Amin Karbasi
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.
****
Optimistic Distributionally Robust Optimization for Nonparametric Likelihood App
roximation
Viet Anh Nguyen, Soroosh Shafieezadeh Abadeh, Man-Chung Yue, Daniel Kuhn, Wolfra
m Wiesemann
The likelihood function is a fundamental component in Bayesian statistics. Howev
er, evaluating the likelihood of an observation is computationally intractable i
n many applications. In this paper, we propose a non-parametric approximation of
 the likelihood that identifies a probability measure which lies in the neighbor
hood of the nominal measure and that maximizes the probability of observing the
given sample point. We show that when the neighborhood is constructed by the Kul
lback-Leibler divergence, by moment conditions or by the Wasserstein distance, t
hen our optimistic likelihood can be determined through the solution of a convex
 optimization problem, and it admits an analytical expression in particular case
s. We also show that the posterior inference problem with our optimistic likelih
ood approximation enjoys strong theoretical performance guarantees, and it perfo
rms competitively in a probabilistic classification task.
****
Don't take it lightly: Phasing optical random projections with unknown operators
Sidharth Gupta, Remi Gribonval, Laurent Daudet, Ivan Dokmani■
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.
****
Visualizing the PHATE of Neural Networks
Scott Gigante, Adam S. Charles, Smita Krishnaswamy, Gal Mishne
Understanding why and how certain neural networks outperform others is key to gu
iding future development of network architectures and optimization methods. To t
his end, we introduce a novel visualization algorithm that reveals the internal
geometry of such networks: Multislice PHATE (M-PHATE), the first method designed
 explicitly to visualize how a neural network's hidden representations of data e
volve throughout the course of training. We demonstrate that our visualization p
rovides intuitive, detailed summaries of the learning dynamics beyond simple glo
bal measures (i.e., validation loss and accuracy), without the need to access va
lidation data. Furthermore, M-PHATE better captures both the dynamics and commun
ity structure of the hidden units as compared to visualization based on standard
 dimensionality reduction methods (e.g., ISOMAP, t-SNE). We demonstrate M-PHATE
with two vignettes: continual learning and generalization. In the former, the M-
PHATE visualizations display the mechanism of "catastrophic forgetting" which is
 a major challenge for learning in task-switching contexts. In the latter, our v
isualizations reveal how increased heterogeneity among hidden units correlates w
ith improved generalization performance. An implementation of M-PHATE, along wit
h scripts to reproduce the figures in this paper, is available at https://github
.com/scottgigante/M-PHATE.
****
Gate Decorator: Global Filter Pruning Method for Accelerating Deep Convolutional
 Neural Networks
Zhonghui You, Kun Yan, Jinmian Ye, Meng Ma, Ping Wang
Filter pruning is one of the most effective ways to accelerate and compress conv
olutional neural networks (CNNs). In this work, we propose a global filter pruni
ng algorithm called Gate Decorator, which transforms a vanilla CNN module by mul

tiplying its output by the channel-wise scaling factors (i.e. gate). When the sc aling factor is set to zero, it is equivalent to removing the corresponding filt er. We use Taylor expansion to estimate the change in the loss function caused b y setting the scaling factor to zero and use the estimation for the global filte r importance ranking. Then we prune the network by removing those unimportant fi lters. After pruning, we merge all the scaling factors into its original module, so no special operations or structures are introduced. Moreover, we propose an iterative pruning framework called Tick-Tock to improve pruning accuracy. The ex tensive experiments demonstrate the effectiveness of our approaches. For example , we achieve the state-of-the-art pruning ratio on ResNet-56 by reducing 70% FLO Ps without noticeable loss in accuracy. For ResNet-50 on ImageNet, our pruned mo del with 40% FLOPs reduction outperforms the baseline model by 0.31% in top-1 ac curacy. Various datasets are used, including CIFAR-10, CIFAR-100, CUB-200, Image Net ILSVRC-12 and PASCAL VOC 2011.

****

Kalman Filter, Sensor Fusion, and Constrained Regression: Equivalences and Insig hts

Maria Jahja, David Farrow, Roni Rosenfeld, Ryan J. Tibshirani

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.

****

Practical Deep Learning with Bayesian Principles

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E. Khan, Anirudh Jain, Runa Es chenhagen, Richard E. Turner, Rio Yokota

Bayesian methods promise to fix many shortcomings of deep learning, but they are impractical and rarely match the performance of standard methods, let alone imp rove them. In this paper, we demonstrate practical training of deep networks wit h natural-gradient variational inference. By applying techniques such as batch n ormalisation, data augmentation, and distributed training, we achieve similar pe rformance in about the same number of epochs as the Adam optimiser, even on larg e datasets such as ImageNet. Importantly, the benefits of Bayesian principles ar e preserved: predictive probabilities are well-calibrated, uncertainties on out- of-distribution data are improved, and continual-learning performance is boosted . This work enables practical deep learning while preserving benefits of Bayesia n principles. A PyTorch implementation is available as a plug-and-play optimiser .

****

Deep Active Learning with a Neural Architecture Search

Yonatan Geifman, Ran El-Yaniv

We consider active learning of deep neural networks. Most active learning works in this context have focused on studying effective querying mechanisms and assum ed that an appropriate network architecture is a priori known for the problem at hand. We challenge this assumption and propose a novel active strategy whereby the learning algorithm searches for effective architectures on the fly, while ac tively learning. We apply our strategy using three known querying techniques (so ftmax response, MC-dropout, and coresets) and show that the proposed approach ov erwhelmingly outperforms active learning using fixed architectures.

****

Quality Aware Generative Adversarial Networks

KANCHARLA PARIMALA, Sumohana Channappayya

Generative Adversarial Networks (GANs) have become a very popular tool for im- plicitly learning high-dimensional probability distributions. Several improvemen ts

have been made to the original GAN formulation to address some of its shortcom- ings like mode collapse, convergence issues, entanglement, poor visual quality e tc.

While a significant effort has been directed towards improving the visual qualit y

of images generated by GANs, it is rather surprising that objective image qualit
y
metrics have neither been employed as cost functions nor as regularizers in GAN
objective functions. In this work, we show how a distance metric that is a varia
nt
of the Structural SIMilarity (SSIM) index (a popular full-reference image qualit
y
assessment algorithm), and a novel quality aware discriminator gradient penalty
function that is inspired by the Natural Image Quality Evaluator (NIQE, a popula
r
no-reference image quality assessment algorithm) can each be used as excellent
regularizers for GAN objective functions. Specifically, we demonstrate state-of-
the-art performance using the Wasserstein GAN gradient penalty (WGAN-GP)
framework over CIFAR-10, STL10 and CelebA datasets.
****
Dual Variational Generation for Low Shot Heterogeneous Face Recognition
Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, Ran He
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.
****
Off-Policy Evaluation via Off-Policy Classification
Alexander Irpan, Kanishka Rao, Konstantinos Bousmalis, Chris Harris, Julian Ibar
z, Sergey Levine
In this work, we consider the problem of model selection for deep reinforcement
learning (RL) in real-world environments. Typically, the performance of deep RL
algorithms is evaluated via on-policy interactions with the target environment.
However, comparing models in a real-world environment for the purposes of early
stopping or hyperparameter tuning is costly and often practically infeasible.
This leads us to examine off-policy policy evaluation (OPE) in such settings.
We focus on OPE of value-based methods, which are of particular interest in deep
 RL with applications like robotics, where off-policy algorithms based on Q-func
tion estimation can often attain better sample complexity than direct policy opt
imization.
Furthermore, existing OPE metrics either rely on a model of the environment, or
the use of importance sampling (IS) to correct for the data being off-policy.
However, for high-dimensional observations, such as images, models of the enviro
nment can be difficult to fit and value-based methods can make IS hard to use or
 even ill-conditioned, especially when dealing with continuous action spaces.
In this paper, we focus on the specific case of MDPs with continuous action spac
es and sparse binary rewards, which is representative of many important real-wor
ld applications. We propose an alternative metric that relies on neither models
nor IS, by framing OPE as a positive-unlabeled (PU) classification problem. We e
xperimentally show that this metric outperforms baselines on a number of tasks.
Most importantly, it can reliably predict the relative performance of different
policies in a number of generalization scenarios, including the transfer to the
real-world of policies trained in simulation for an image-based robotic manipula
tion task.
****
Variational Temporal Abstraction
Taesup Kim, Sungjin Ahn, Yoshua Bengio
We introduce a variational approach to learning and inference of temporally hier
archical structure and representation for sequential data. We propose the Variat
ional Temporal Abstraction (VTA), a hierarchical recurrent state space model tha
t can infer the latent temporal structure and thus perform the stochastic state
transition hierarchically. We also propose to apply this model to implement the
jumpy imagination ability in imagination-augmented agent-learning in order to im
prove the efficiency of the imagination. In experiments, we demonstrate that our
 proposed method can model 2D and 3D visual sequence datasets with interpretable

temporal structure discovery and that its application to jumpy imagination enables more efficient agent-learning in a 3D navigation task.
****
Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations
Vincent Sitzmann, Michael Zollhoefer, Gordon Wetzstein
Unsupervised learning with generative models has the potential of discovering rich representations of 3D scenes. While geometric deep learning has explored 3D-structure-aware representations of scene geometry, these models typically require explicit 3D supervision. Emerging neural scene representations can be trained only with posed 2D images, but existing methods ignore the three-dimensional structure of scenes. We propose Scene Representation Networks (SRNs), a continuous, 3D-structure-aware scene representation that encodes both geometry and appearance. SRNs represent scenes as continuous functions that map world coordinates to a feature representation of local scene properties. By formulating the image formation as a differentiable ray-marching algorithm, SRNs can be trained end-to-end from only 2D images and their camera poses, without access to depth or shape. This formulation naturally generalizes across scenes, learning powerful geometry and appearance priors in the process. We demonstrate the potential of SRNs by evaluating them for novel view synthesis, few-shot reconstruction, joint shape and appearance interpolation, and unsupervised discovery of a non-rigid face model.
****
Control What You Can: Intrinsically Motivated Task-Planning Agent
Sebastian Blaes, Marin Vlastelica Pogan■i■, Jiajie Zhu, Georg Martius
We present a novel intrinsically motivated agent that learns how to control the environment in a sample efficient manner, that is with as few environment interactions as possible, by optimizing learning progress. It learns what can be controlled, how to allocate time and attention as well as the relations between objects using surprise-based motivation. The effectiveness of our method is demonstrated in a synthetic and robotic manipulation environment yielding considerably improved performance and smaller sample complexity compared to an intrinsically motivated, non-hierarchical and state-of-the-art hierarchical baseline. In a nutshell, our work combines several task-level planning agent structures (backtracking search on task-graph, probabilistic road-maps, allocation of search efforts) with intrinsic motivation to achieve learning from scratch.
****
Momentum-Based Variance Reduction in Non-Convex SGD
Ashok Cutkosky, Francesco Orabona
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 Self-Defense for Cycle-Consistent GANs
Dina Bashkirova, Ben Usman, Kate Saenko
The goal of unsupervised image-to-image translation is to map images from one domain to another without the ground truth correspondence between the two domains. State-of-art methods learn the correspondence using large numbers of unpaired examples from both domains and are based on generative adversarial networks. In order to preserve the semantics of the input image, the adversarial objective is usually combined with a cycle-consistency loss that penalizes incorrect reconstruction of the input image from the translated one. However, if the target mapping is many-to-one, e.g. aerial photos to maps, such a restriction forces the generator to hide information in low-amplitude structured noise that is undetectable by human eye or by the discriminator. In this paper, we show how such self-attacking behavior of unsupervised translation methods affects their performance and provide two defense techniques. We perform a quantitative evaluation of the proposed techniques and show that making the translation model more robust to the self-adversarial attack increases its generation quality and reconstruction reliability and makes the model less sensitive to low-amplitude perturbations. Our

project page can be found at ai.bu.edu/selfadv.
****

Ultrametric Fitting by Gradient Descent
Giovanni Chierchia, Benjamin Perret
We study the problem of fitting an ultrametric distance to a dissimilarity graph in the context of hierarchical cluster analysis. Standard hierarchical clustering methods are specified procedurally, rather than in terms of the cost function to be optimized. We aim to overcome this limitation by presenting a general optimization framework for ultrametric fitting. Our approach consists of modeling the latter as a constrained optimization problem over the continuous space of ultrametrics. So doing, we can leverage the simple, yet effective, idea of replacing the ultrametric constraint with a min-max operation injected directly into the cost function. The proposed reformulation leads to an unconstrained optimization problem that can be efficiently solved by gradient descent methods. The flexibility of our framework allows us to investigate several cost functions, following the classic paradigm of combining a data fidelity term with a regularization. While we provide no theoretical guarantee to find the global optimum, the numerical results obtained over a number of synthetic and real datasets demonstrate the good performance of our approach with respect to state-of-the-art agglomerative algorithms. This makes us believe that the proposed framework sheds new light on the way to design a new generation of hierarchical clustering methods. Our code is made publicly available
at https://github.com/PerretB/ultrametric-fitting.
****

Expressive power of tensor-network factorizations for probabilistic modeling
Ivan Glasser, Ryan Sweke, Nicola Pancotti, Jens Eisert, Ignacio Cirac
Tensor-network techniques have recently proven useful in machine learning, both as a tool for the formulation of new learning algorithms and for enhancing the mathematical understanding of existing methods. Inspired by these developments, and the natural correspondence between tensor networks and probabilistic graphical models, we provide a rigorous analysis of the expressive power of various tensor-network factorizations of discrete multivariate probability distributions. These factorizations include non-negative tensor-trains/MPS, which are in correspondence with hidden Markov models, and Born machines, which are naturally related to the probabilistic interpretation of quantum circuits. When used to model probability distributions, they exhibit tractable likelihoods and admit efficient learning algorithms. Interestingly, we prove that there exist probability distributions for which there are unbounded separations between the resource requirements of some of these tensor-network factorizations. Of particular interest, using complex instead of real tensors can lead to an arbitrarily large reduction in the number of parameters of the network. Additionally, we introduce locally purified states (LPS), a new factorization inspired by techniques for the simulation of quantum systems, with provably better expressive power than all other representations considered. The ramifications of this result are explored through numerical experiments.
****

PerspectiveNet: 3D Object Detection from a Single RGB Image via Perspective Points
Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, Song-Chun Zhu
Detecting 3D objects from a single RGB image is intrinsically ambiguous, thus requiring appropriate prior knowledge and intermediate representations as constraints to reduce the uncertainties and improve the consistencies between the 2D image plane and the 3D world coordinate. To address this challenge, we propose to adopt perspective points as a new intermediate representation for 3D object detection, defined as the 2D projections of local Manhattan 3D keypoints to locate an object; these perspective points satisfy geometric constraints imposed by the perspective projection. We further devise PerspectiveNet, an end-to-end trainable model that simultaneously detects the 2D bounding box, 2D perspective points, and 3D object bounding box for each object from a single RGB image. PerspectiveNet yields three unique advantages: (i) 3D object bounding boxes are estimated bas

ed on perspective points, bridging the gap between 2D and 3D bounding boxes without the need of category-specific 3D shape priors. (ii) It predicts the perspective points by a template-based method, and a perspective loss is formulated to maintain the perspective constraints. (iii) It maintains the consistency between the 2D perspective points and 3D bounding boxes via a differentiable projective function. Experiments on SUN RGB-D dataset show that the proposed method significantly outperforms existing RGB-based approaches for 3D object detection.
****

Landmark Ordinal Embedding

Nikhil Ghosh, Yuxin Chen, Yisong Yue

In this paper, we aim to learn a low-dimensional Euclidean representation from a set of constraints of the form “item j is closer to item i than item k”. Existing approaches for this “ordinal embedding” problem require expensive optimization procedures, which cannot scale to handle increasingly larger datasets. To address this issue, we propose a landmark-based strategy, which we call Landmark Ordinal Embedding (LOE). Our approach trades off statistical efficiency for computational efficiency by exploiting the low-dimensionality of the latent embedding. We derive bounds establishing the statistical consistency of LOE under the popular Bradley- Terry-Luce noise model. Through a rigorous analysis of the computational complexity, we show that LOE is significantly more efficient than conventional ordinal embedding approaches as the number of items grows. We validate these characterizations empirically on both synthetic and real datasets. We also present a practical approach that achieves the “best of both worlds”, by using LOE to warm-start existing methods that are more statistically efficient but computationally expensive.
****

On the Value of Target Data in Transfer Learning

Steve Hanneke, Samory Kpotufe

We aim to understand the value of additional labeled or unlabeled target data in transfer learning, for any given amount of source data; this is motivated by practical questions around minimizing sampling costs, whereby, target data is usually harder or costlier to acquire than source data, but can yield better accuracy.
****

Machine Teaching of Active Sequential Learners

Tomi Peltola, Mustafa Mert Çelikok, Pedram Daee, Samuel Kaski

Machine teaching addresses the problem of finding the best training data that can guide a learning algorithm to a target model with minimal effort. In conventional settings, a teacher provides data that are consistent with the true data distribution. However, for sequential learners which actively choose their queries, such as multi-armed bandits and active learners, the teacher can only provide responses to the learner’s queries, not design the full data. In this setting, consistent teachers can be sub-optimal for finite horizons. We formulate this sequential teaching problem, which current techniques in machine teaching do not address, as a Markov decision process, with the dynamics nesting a model of the learner and the actions being the teacher's responses. Furthermore, we address the complementary problem of learning from a teacher that plans: to recognise the teaching intent of the responses, the learner is endowed with a model of the teacher. We test the formulation with multi-armed bandit learners in simulated experiments and a user study. The results show that learning is improved by (i) planning teaching and (ii) the learner having a model of the teacher. The approach gives tools to taking into account strategic (planning) behaviour of users of interactive intelligent systems, such as recommendation engines, by considering them as boundedly optimal teachers.
****

Beyond Confidence Regions: Tight Bayesian Ambiguity Sets for Robust MDPs

Marek Petrik, Reazul Hasan Russel

Robust MDPs (RMDPs) can be used to compute policies with provable worst-case guarantees in reinforcement learning. The quality and robustness of an RMDP solution are determined by the ambiguity set---the set of plausible transition probabil

ities---which is usually constructed as a multi-dimensional confidence region. Existing methods construct ambiguity sets as confidence regions using concentration inequalities which leads to overly conservative solutions. This paper proposes a new paradigm that can achieve better solutions with the same robustness guarantees without using confidence regions as ambiguity sets. To incorporate prior knowledge, our algorithms optimize the size and position of ambiguity sets using Bayesian inference. Our theoretical analysis shows the safety of the proposed method, and the empirical results demonstrate its practical promise.
****

A General Theory of Equivariant CNNs on Homogeneous Spaces
Taco S. Cohen, Mario Geiger, Maurice Weiler
We present a general theory of Group equivariant Convolutional Neural Networks (G-CNNs) on homogeneous spaces such as Euclidean space and the sphere. Feature maps in these networks represent fields on a homogeneous base space, and layers are equivariant maps between spaces of fields. The theory enables a systematic classification of all existing G-CNNs in terms of their symmetry group, base space, and field type. We also answer a fundamental question: what is the most general kind of equivariant linear map between feature spaces (fields) of given types? We show that such maps correspond one-to-one with generalized convolutions with an equivariant kernel, and characterize the space of such kernels.
****

Spatial-Aware Feature Aggregation for Image based Cross-View Geo-Localization
Yujiao Shi, Liu Liu, Xin Yu, Hongdong Li
In this paper, we develop a new deep network to explicitly address these inherent differences between ground and aerial views. We observe there exist some approximate domain correspondences between ground and aerial images. Specifically, pixels lying on the same azimuth direction in an aerial image approximately correspond to a vertical image column in the ground view image. Thus, we propose a two-step approach to exploit this prior knowledge. The first step is to apply a regular polar transform to warp an aerial image such that its domain is closer to that of a ground-view panorama. Note that polar transform as a pure geometric transformation is agnostic to scene content, hence cannot bring the two domains into full alignment. Then, we add a subsequent spatial-attention mechanism which further brings corresponding deep features closer in the embedding space. To improve the robustness of feature representation, we introduce a feature aggregation strategy via learning multiple spatial embeddings. By the above two-step approach, we achieve more discriminative deep representations, facilitating cross-view Geo-localization more accurate. Our experiments on standard benchmark datasets show significant performance boosting, achieving more than doubled recall rate compared with the previous state of the art.
****

Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification
Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, Massimiliano Pontil
We study the problem of fair binary classification using the notion of Equal Opportunity.
It requires the true positive rate to distribute equally across the sensitive groups.
Within this setting we show that the fair optimal classifier is obtained by recalibrating the Bayes classifier by a group-dependent threshold. We provide a constructive expression for the threshold.
This result motivates us to devise a plug-in classification procedure based on both unlabeled and labeled datasets.
While the latter is used to learn the output conditional probability, the former is used for calibration.
The overall procedure can be computed in polynomial time and it is shown to be statistically consistent both in terms of the classification error and fairness measure. Finally, we present numerical experiments which indicate that our method is often superior or competitive with the state-of-the-art methods on benchmark datasets.

****
Tight Dimensionality Reduction for Sketching Low Degree Polynomial Kernels
Michela Meister, Tamas Sarlos, 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.
****
Minimum Stein Discrepancy Estimators
Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, Lester Mackey
When maximum likelihood estimation is infeasible, one often turns to score matching, contrastive divergence, or minimum probability flow to obtain tractable parameter estimates. We provide a unifying perspective of these techniques as minimum Stein discrepancy estimators, and use this lens to design new diffusion kernel Stein discrepancy (DKSD) and diffusion score matching (DSM) estimators with complementary strengths. We establish the consistency, asymptotic normality, and robustness of DKSD and DSM estimators, then derive stochastic Riemannian gradient descent algorithms for their efficient optimisation. The main strength of our methodology is its flexibility, which allows us to design estimators with desirable properties for specific models at hand by carefully selecting a Stein discrepancy.
We illustrate this advantage for several challenging problems for score matching, such as non-smooth, heavy-tailed or light-tailed densities.
****
Provably Powerful Graph Networks
Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, Yaron Lipman
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.
****
Regularized Anderson Acceleration for Off-Policy Deep Reinforcement Learning
Wenjie Shi, Shiji Song, Hui Wu, Ya-Chu Hsu, Cheng Wu, Gao Huang
Model-free deep reinforcement learning (RL) algorithms have been widely used for a range of complex control tasks. However, slow convergence and sample inefficiency remain challenging problems in RL, especially when handling continuous and high-dimensional state spaces. To tackle this problem, we propose a general acceleration method for model-free, off-policy deep RL algorithms by drawing the idea underlying regularized Anderson acceleration (RAA), which is an effective approach to accelerating the solving of fixed point problems with perturbations. Specifically, we first explain how policy iteration can be applied directly with Anderson acceleration. Then we extend RAA to the case of deep RL by introducing a regularization term to control the impact of perturbation induced by function approximation errors. We further propose two strategies, i.e., progressive update and adaptive restart, to enhance the performance. The effectiveness of our method is evaluated on a variety of benchmark tasks, including Atari 2600 and MuJoCo. Experimental results show that our approach substantially improves both the learning speed and final performance of state-of-the-art deep RL algorithms.
****
Kernel Stein Tests for Multiple Model Comparison
Jen Ning Lim, Makoto Yamada, Bernhard Schölkopf, Wittawat Jitkrittum
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 can be manipulated and geometry is to blame
Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, Pan Kessel

Explanation methods aim to make neural networks more trustworthy and interpretab
le. In this paper, we demonstrate a property of explanation methods which is dis
concerting for both of these purposes. Namely, we show that explanations can be
manipulated arbitrarily by applying visually hardly perceptible perturbations to
 the input that keep the network's output approximately constant. We establish t
heoretically that this phenomenon can be related to certain geometrical properti
es of neural networks. This allows us to derive an upper bound on the susceptibi
lity of explanations to manipulations. Based on this result, we propose effectiv
e mechanisms to enhance the robustness of explanations.
****
Input-Cell Attention Reduces Vanishing Saliency of Recurrent Neural Networks
Aya Abdelsalam Ismail, Mohamed Gunady, Luiz Pessoa, Hector Corrada Bravo, Soheil
 Feizi
Recent efforts to improve the interpretability of deep neural networks use salie
ncy to characterize the importance of input features to predictions made by mode
ls. Work on interpretability using saliency-based methods on Recurrent Neural Ne
tworks (RNNs) has mostly targeted language tasks, and their applicability to tim
e series data is less understood. In this work we analyze saliency-based methods
 for RNNs, both classical and gated cell architectures. We show that RNN salienc
y vanishes over time, biasing detection of salient features only to later time s
teps and are, therefore, incapable of reliably detecting important features at a
rbitrary time intervals. To address this vanishing saliency problem, we propose
a novel RNN cell structure (input-cell attention), which can extend any RNN cell
 architecture. At each time step, instead of only looking at the current input v
ector, input-cell attention uses a fixed-size matrix embedding, each row of the
matrix attending to different inputs from current or previous time steps. Using
 synthetic data, we show that the saliency map produced by the input-cell attent
ion RNN is able to faithfully detect important features regardless of their occu
rrence in time. We also apply the input-cell attention RNN on a neuroscience tas
k analyzing functional Magnetic Resonance Imaging (fMRI) data for human subjects
 performing a variety of tasks. In this case, we use saliency to characterize br
ain regions (input features) for which activity is important to distinguish betw
een tasks. We show that standard RNN architectures are only capable of detecting
 important brain regions in the last few time steps of the fMRI data, while the
input-cell attention model is able to detect important brain region activity acr
oss time without latter time step biases.
****
Paradoxes in Fair Machine Learning
Paul Goelz, Anson Kahng, Ariel D. Procaccia
Equalized odds is a statistical notion of fairness in machine learning that ensu
res that classification algorithms do not discriminate against protected groups.
 We extend equalized odds to the setting of cardinality-constrained fair classif
ication, where we have a bounded amount of a resource to distribute. This settin
g coincides with classic fair division problems, which allows us to apply concep
ts from that literature in parallel to equalized odds. In particular, we conside
r the axioms of resource monotonicity, consistency, and population monotonicity,
 all three of which relate different allocation instances to prevent paradoxes.
Using a geometric characterization of equalized odds, we examine the compatibili
ty of equalized odds with these axioms. We empirically evaluate the cost of allo
cation rules that satisfy both equalized odds and axioms of fair division on a d
ataset of FICO credit scores.
****
Learning Conditional Deformable Templates with Convolutional Networks
Adrian Dalca, Marianne Rakic, John Guttag, Mert Sabuncu
We develop a learning framework for building deformable templates, which play a
fundamental role in many image analysis and computational anatomy tasks. Convent
ional methods for template creation and image alignment to the template have und
ergone decades of rich technical development. In these frameworks, templates are
 constructed using an iterative process of template estimation and alignment, wh
ich is often computationally very expensive. Due in part to this shortcoming, mo

st methods compute a single template for the entire population of images, or a few templates for specific sub-groups of the data. In this work, we present a probabilistic model and efficient learning strategy that yields either universal or \textit{conditional} templates, jointly with a neural network that provides efficient alignment of the images to these templates. We demonstrate the usefulness of this method on a variety of domains, with a special focus on neuroimaging. This is particularly useful for clinical applications where a pre-existing template does not exist, or creating a new one with traditional methods can be prohibitively expensive. Our code and atlases are available online as part of the VoxelMorph library at http://voxelmorph.csail.mit.edu.
****
Volumetric Correspondence Networks for Optical Flow
Gengshan Yang, Deva Ramanan
Many classic tasks in vision -- such as the estimation of optical flow or stereo disparities -- can be cast as dense correspondence matching. Well-known techniques for doing so make use of a cost volume, typically a 4D tensor of match costs between all pixels in a 2D image and their potential matches in a 2D search window. State-of-the-art (SOTA) deep networks for flow/stereo make use of such volumetric representations as internal layers. However, such layers require significant amounts of memory and compute, making them cumbersome to use in practice. As a result, SOTA networks also employ various heuristics designed to limit volumetric processing, leading to limited accuracy and overfitting. Instead, we introduce several simple modifications that dramatically simplify the use of volumetric layers - (1) volumetric encoder-decoder architectures that efficiently capture large receptive fields, (2) multi-channel cost volumes that capture multi-dimensional notions of pixel similarities, and finally, (3) separable volumetric filtering that significantly reduces computation and parameters while preserving accuracy. Our innovations dramatically improve accuracy over SOTA on standard benchmarks while being significantly easier to work with - training converges in 10X fewer iterations, and most importantly, our networks generalize across correspondence tasks. On-the-fly adaptation of search windows allows us to repurpose optical flow networks for stereo (and vice versa), and can also be used to implement adaptive networks that increase search window sizes on-demand.
****
Variance Reduction in Bipartite Experiments through Correlation Clustering
Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, Vahab Mirrokni
Causal inference in randomized experiments typically assumes that the units of randomization and the units of analysis are one and the same. In some applications, however, these two roles are played by distinct entities linked by a bipartite graph. The key challenge in such bipartite settings is how to avoid interference bias, which would typically arise if we simply randomized the treatment at the level of analysis units. One effective way of minimizing interference bias in standard experiments is through cluster randomization, but this design has not been studied in the bipartite setting where conventional clustering schemes can lead to poorly powered experiments. This paper introduces a novel clustering objective and a corresponding algorithm that partitions a bipartite graph so as to maximize the statistical power of a bipartite experiment on that graph. Whereas previous work relied on balanced partitioning, our formulation suggests the use of a correlation clustering objective. We use a publicly-available graph of Amazon user-item reviews to validate our solution and illustrate how it substantially increases the statistical power in bipartite experiments.
****
Attribution-Based Confidence Metric For Deep Neural Networks
Susmit Jha, Sunny Raj, Steven Fernandes, Sumit K. Jha, Somesh Jha, Brian Jalaian, Gunjan Verma, Ananthram Swami
We propose a novel confidence metric, namely, attribution-based confidence (ABC) for deep neural networks (DNNs). ABC metric characterizes whether the output of a DNN on an input can be trusted. DNNs are known to be brittle on inputs outside the training distribution and are, hence, susceptible to adversarial attacks. This fragility is compounded by a lack of effectively computable measures of mo

del confidence that correlate well with the accuracy of DNNs. These factors have impeded the adoption of DNNs in high-assurance systems. The proposed ABC metric addresses these challenges. It does not require access to the training data, the use of ensembles, or the need to train a calibration model on a held-out validation set. Hence, the new metric is usable even when only a trained model is available for inference. We mathematically motivate the proposed metric and evaluate its effectiveness with two sets of experiments. First, we study the change in accuracy and the associated confidence over out-of-distribution inputs. Second, we consider several digital and physically realizable attacks such as FGSM, CW, DeepFool, PGD, and adversarial patch generation methods. The ABC metric is low on out-of-distribution data and adversarial examples, where the accuracy of the model is also low. These experiments demonstrate the effectiveness of the ABC metric to make DNNs more trustworthy and resilient.
****

Are Disentangled Representations Helpful for Abstract Visual Reasoning?

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, Olivier Bachem

A disentangled representation encodes information about the salient factors of variation in the data independently. Although it is often argued that this representational format is useful in learning to solve many real-world down-stream tasks, there is little empirical evidence that supports this claim. In this paper, we conduct a large-scale study that investigates whether disentangled representations are more suitable for abstract reasoning tasks. Using two new tasks similar to Raven's Progressive Matrices, we evaluate the usefulness of the representations learned by 360 state-of-the-art unsupervised disentanglement models. Based on these representations, we train 3600 abstract reasoning models and observe that disentangled representations do in fact lead to better down-stream performance. In particular, they enable quicker learning using fewer samples.
****

RSN: Randomized Subspace Newton

Robert Gower, Dmitry Kovalev, Felix Lieder, Peter Richtarik

We develop a randomized Newton method capable of solving learning problems with huge dimensional feature spaces, which is a common setting in applications such as medical imaging, genomics and seismology. Our method leverages randomized sketching in a new way, by finding the Newton direction constrained to the space spanned by a random sketch. We develop a simple global linear convergence theory that holds for practically all sketching techniques, which gives the practitioners the freedom to design custom sketching approaches suitable for particular applications. We perform numerical experiments which demonstrate the efficiency of our method as compared to accelerated gradient descent and the full Newton method. Our method can be seen as a refinement and a randomized extension of the results of Karimireddy, Stich, and Jaggi (2019).
****

Beyond Alternating Updates for Matrix Factorization with Inertial Bregman Proximal Gradient Algorithms

Mahesh Chandra Mukkamala, Peter Ochs

Matrix Factorization is a popular non-convex optimization problem, for which alternating minimization schemes are mostly used. They usually suffer from the major drawback that the solution is biased towards one of the optimization variables. A remedy is non-alternating schemes. However, due to a lack of Lipschitz continuity of the gradient in matrix factorization problems, convergence cannot be guaranteed. A recently developed approach relies on the concept of Bregman distances, which generalizes the standard Euclidean distance. We exploit this theory by proposing a novel Bregman distance for matrix factorization problems, which, at the same time, allows for simple/closed form update steps. Therefore, for non-alternating schemes, such as the recently introduced Bregman Proximal Gradient (BPG) method and an inertial variant Convex--Concave Inertial BPG (CoCaIn BPG), convergence of the whole sequence to a stationary point is proved for Matrix Factorization. In several experiments, we observe a superior performance of our non-alternating schemes in terms of speed and objective value at the limit point.
****

Integrating Bayesian and Discriminative Sparse Kernel Machines for Multi-class Active Learning
Weishi Shi, Qi Yu
We propose a novel active learning (AL) model that integrates Bayesian and discr iminative kernel machines for fast and accurate multi-class data sampling. By jo ining a sparse Bayesian model and a maximum margin machine under a unified kerne l machine committee (KMC), the proposed model is able to identify a small number of data samples that best represent the overall data space while accurately cap turing the decision boundaries. The integration is conducted using the maximum entropy discrimination framework, resulting in a joint objective function that c ontains generalized entropy as a regularizer. Such a property allows the propose d AL model to choose data samples that more effectively handle non-separable cla ssification problems. Parameter learning is achieved through a principled optimi zation framework that leverages convex duality and sparse structure of KMC to e fficiently optimize the joint objective function. Key model parameters are used to design a novel sampling function to choose data samples that can simultaneou sly improve multiple decision boundaries, making it an effective sampler for pro blems with a large number of classes. Experiments conducted over both synthetic and real data and comparison with competitive AL methods demonstrate the effecti veness of the proposed model.
****
Towards Explaining the Regularization Effect of Initial Large Learning Rate in T raining Neural Networks
Yuanzhi Li, Colin Wei, Tengyu Ma
Stochastic gradient descent with a large initial learning rate is widely used fo r training modern neural net architectures. Although a small initial learning ra te allows for faster training and better test performance initially, the large l earning rate achieves better generalization soon after the learning rate is anne aled. Towards explaining this phenomenon, we devise a setting in which we can pr ove that a two layer network trained with large initial learning rate and anneal ing provably generalizes better than the same network trained with a small learn ing rate from the start. The key insight in our analysis is that the order of le arning different types of patterns is crucial: because the small learning rate m odel first memorizes low-noise, hard-to-fit patterns, it generalizes worse on ha rd-to-generalize, easier-to-fit patterns than its large learning rate counterpar t. This concept translates to a larger-scale setting: we demonstrate that one ca n add a small patch to CIFAR-10 images that is immediately memorizable by a mode l with small initial learning rate, but ignored by the model with large learning rate until after annealing. Our experiments show that this causes the small lea rning rate model's accuracy on unmodified images to suffer, as it relies too muc h on the patch early on.
****
An Algorithm to Learn Polytree Networks with Hidden Nodes
Firoozeh Sepehr, Donatello Materassi
Ancestral graphs are a prevalent mathematical tool to take into account latent (hidden) variables in a probabilistic graphical model. In ancestral graph represe ntations, the nodes are only the observed (manifest) variables and the notion of m-separation fully characterizes the conditional independence relations among s uch variables, bypassing the need to explicitly consider latent variables. Howev er, ancestral graph models do not necessarily represent the actual causal struct ure of the model, and do not contain information about, for example, the precise number and location of the hidden variables. Being able to detect the presence of latent variables while also inferring their precise location within the actua l causal structure model is a more challenging task that provides more informati on about the actual causal relationships among all the model variables, includin g the latent ones. In this article, we develop an algorithm to exactly recover g raphical models of random variables with underlying polytree structures when the latent nodes satisfy specific degree conditions. Therefore, this article propos es an approach for the full identification of hidden variables in a polytree. We also show that the algorithm is complete in the sense that when such degree con

ditions are not met, there exists another polytree with fewer number of latent nodes satisfying the degree conditions and entailing the same independence relations among the observed variables, making it indistinguishable from the actual polytree.
****
Provable Gradient Variance Guarantees for Black-Box Variational Inference
Justin Domke
Recent variational inference methods use stochastic gradient estimators whose variance is not well understood. Theoretical guarantees for these estimators are important to understand when these methods will or will not work. This paper gives bounds for the common “reparameterization” estimators when the target is smooth and the variational family is a location-scale distribution. These bounds are unimprovable and thus provide the best possible guarantees under the stated assumptions.
****
LiteEval: A Coarse-to-Fine Framework for Resource Efficient Video Recognition
Zuxuan Wu, Caiming Xiong, Yu-Gang Jiang, Larry S. Davis
This paper presents LiteEval, a simple yet effective coarse-to-fine framework for resource efficient video recognition, suitable for both online and offline scenarios. Exploiting decent yet computationally efficient features derived at a coarse scale with a lightweight CNN model, LiteEval dynamically decides on-the-fly whether to compute more powerful features for incoming video frames at a finer scale to obtain more details. This is achieved by a coarse LSTM and a fine LSTM operating cooperatively, as well as a conditional gating module to learn when to allocate more computation. Extensive experiments are conducted on two large-scale video benchmarks, FCVID and ActivityNet, and the results demonstrate LiteEval requires substantially less computation while offering excellent classification accuracy for both online and offline predictions.
****
Multi-marginal Wasserstein GAN
Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, Mingkui Tan
Multiple marginal matching problem aims at learning mappings to match a source domain to multiple target domains and it has attracted great attention in many applications, such as multi-domain image translation. However, addressing this problem has two critical challenges: (i) Measuring the multi-marginal distance among different domains is very intractable; (ii) It is very difficult to exploit cross-domain correlations to match the target domain distributions. In this paper, we propose a novel Multi-marginal Wasserstein GAN (MWGAN) to minimize Wasserstein distance among domains. Specifically, with the help of multi-marginal optimal transport theory, we develop a new adversarial objective function with inner- and inter-domain constraints to exploit cross-domain correlations. Moreover, we theoretically analyze the generalization performance of MWGAN, and empirically evaluate it on the balanced and imbalanced translation tasks. Extensive experiments on toy and real-world datasets demonstrate the effectiveness of MWGAN.
****
PyTorch: An Imperative Style, High-Performance Deep Learning Library
Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala
Deep learning frameworks have often focused on either usability or speed, but not both. PyTorch is a machine learning library that shows that these two goals are in fact compatible: it was designed from first principles to support an imperative and Pythonic programming style that supports code as a model, makes debugging easy and is consistent with other popular scientific computing libraries, while remaining efficient and supporting hardware accelerators such as GPUs.
In this paper, we detail the principles that drove the implementation of PyTorch and how they are reflected in its architecture. We emphasize that every aspect of PyTorch is a regular Python program under the full control of its user. We also explain how the careful and pragmatic implementation of the key components of

its runtime enables them to work together to achieve compelling performance. We demonstrate the efficiency of individual subsystems, as well as the overall speed of PyTorch on several commonly used benchmarks.
****

Learning to Infer Implicit Surfaces without 3D Supervision
Shichen Liu, Shunsuke Saito, Weikai Chen, Hao Li
Recent advances in 3D deep learning have shown that it is possible to train highly effective deep models for 3D shape generation, directly from 2D images. This is particularly interesting since the availability of 3D models is still limited compared to the massive amount of accessible 2D images, which is invaluable for training. The representation of 3D surfaces itself is a key factor for the quality and resolution of the 3D output. While explicit representations, such as point clouds and voxels, can span a wide range of shape variations, their resolutions are often limited. Mesh-based representations are more efficient but are limited by their ability to handle varying topologies. Implicit surfaces, however, can robustly handle complex shapes, topologies, and also provide flexible resolution control. We address the fundamental problem of learning implicit surfaces for shape inference without the need of 3D supervision. Despite their advantages, it remains nontrivial to (1) formulate a differentiable connection between implicit surfaces and their 2D renderings, which is needed for image-based supervision; and (2) ensure precise geometric properties and control, such as local smoothness. In particular, sampling implicit surfaces densely is also known to be a computationally demanding and very slow operation. To this end, we propose a novel ray-based field probing technique for efficient image-to-field supervision, as well as a general geometric regularizer for implicit surfaces, which provides natural shape priors in unconstrained regions. We demonstrate the effectiveness of our framework on the task of single-view image-based 3D shape digitization and show how we outperform state-of-the-art techniques both quantitatively and qualitatively.
****

On Sample Complexity Upper and Lower Bounds for Exact Ranking from Noisy Comparisons
Wenbo Ren, Jia (Kevin) Liu, Ness Shroff
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.
****

Are Labels Required for Improving Adversarial Robustness?
Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, Pushmeet Kohli
Recent work has uncovered the interesting (and somewhat surprising) finding that training models to be invariant to adversarial perturbations requires substantially larger datasets than those required for standard classification. This result is a key hurdle in the deployment of robust machine learning models in many real world applications where labeled data is expensive. Our main insight is that unlabeled data can be a competitive alternative to labeled data for training adversarially robust models. Theoretically, we show that in a simple statistical setting, the sample complexity for learning an adversarially robust model from unlabeled data matches the fully supervised case up to constant factors. On standard datasets like CIFAR- 10, a simple Unsupervised Adversarial Training (UAT) approach using unlabeled data improves robust accuracy by 21.7% over using 4K supervised examples alone, and captures over 95% of the improvement from the same number of labeled examples. Finally, we report an improvement of 4% over the previous state-of-the- art on CIFAR-10 against the strongest known attack by using additional unlabeled data from the uncurated 80 Million Tiny Images dataset. This demonstrates that our finding extends as well to the more realistic case where unlabeled data is also uncurated, therefore opening a new avenue for improving adversarial training.
****

NAT: Neural Architecture Transformer for Accurate and Compact Architectures
Yong Guo, Yin Zheng, Mingkui Tan, Qi Chen, Jian Chen, Peilin Zhao, Junzhou Huang
Designing effective architectures is one of the key factors behind the success o
f deep neural networks. Existing deep architectures are either manually designed
 or automatically searched by some Neural Architecture Search (NAS) methods. How
ever, even a well-searched architecture may still contain many non-significant o
r redundant modules or operations (e.g., convolution or pooling), which may not
only incur substantial memory consumption and computation cost but also deterior
ate the performance. Thus, it is necessary to optimize the operations inside an
architecture to improve the performance without introducing extra computation co
st. Unfortunately, such a constrained optimization problem is NP-hard. To make t
he problem feasible, we cast the optimization problem into a Markov decision pro
cess (MDP) and seek to learn a Neural Architecture Transformer (NAT) to replace
the redundant operations with the more computationally efficient ones (e.g., ski
p connection or directly removing the connection). Based on MDP, we learn NAT by
 exploiting reinforcement learning to obtain the optimization policies w.r.t. di
fferent architectures. To verify the effectiveness of the proposed strategies, w
e apply NAT on both hand-crafted architectures and NAS based architectures. Exte
nsive experiments on two benchmark datasets, i.e., CIFAR-10 and ImageNet, demons
trate that the transformed architecture by NAT significantly outperforms both it
s original form and those architectures optimized by existing methods.
****
Learning to Self-Train for Semi-Supervised Few-Shot Classification
Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, Bernt
Schiele
Few-shot classification (FSC) is challenging due to the scarcity of labeled trai
ning data (e.g. only one labeled data point per class). Meta-learning has shown
to achieve promising results by learning to initialize a classification model fo
r FSC. In this paper we propose a novel semi-supervised meta-learning method cal
led learning to self-train (LST) that leverages unlabeled data and specifically
meta-learns how to cherry-pick and label such unsupervised data to further impro
ve performance. To this end, we train the LST model through a large number of se
mi-supervised few-shot tasks. On each task, we train a few-shot model to predict
 pseudo labels for unlabeled data, and then iterate the self-training steps on l
abeled and pseudo-labeled data with each step followed by fine-tuning. We additi
onally learn a soft weighting network (SWN) to optimize the self-training weight
s of pseudo labels so that better ones can contribute more to gradient descent o
ptimization. We evaluate our LST method on two ImageNet benchmarks for semi-supe
rvised few-shot classification and achieve large improvements over the state-of-
the-art.
****
Stochastic Frank-Wolfe for Composite Convex Minimization
Francesco Locatello, Alp Yurtsever, Olivier Fercoq, Volkan Cevher
A broad class of convex optimization problems can be formulated as a semidefinit
e program (SDP), minimization of a convex function over the positive-semidefinit
e cone subject to some affine constraints. The majority of classical SDP solvers
 are designed for the deterministic setting where problem data is readily availa
ble. In this setting, generalized conditional gradient methods (aka Frank-Wolfe-
type methods) provide scalable solutions by leveraging the so-called linear mini
mization oracle instead of the projection onto the semidefinite cone. Most probl
ems in machine learning and modern engineering applications, however, contain so
me degree of stochasticity. In this work, we propose the first conditional-gradi
ent-type method for solving stochastic optimization problems under affine constr
aints. Our method guarantees O(k^{-1/3}) convergence rate in expectation on the
objective residual and O(k^{-5/12}) on the feasibility gap.
****
Modeling Dynamic Functional Connectivity with Latent Factor Gaussian Processes
Lingge Li, Dustin Pluta, Babak Shahbaba, Norbert Fortin, Hernando Ombao, Pierre
Baldi
Dynamic functional connectivity, as measured by the time-varying covariance of n

eurological signals, is believed to play an important role in many aspects of cognition. While many methods have been proposed, reliably establishing the presence and characteristics of brain connectivity is challenging due to the high dimensionality and noisiness of neuroimaging data. We present a latent factor Gaussian process model which addresses these challenges by learning a parsimonious representation of connectivity dynamics. The proposed model naturally allows for inference and visualization of the time-varying connectivity. As an illustration of the scientific utility of the model, application to a data set of rat local field potential activity recorded during a complex non-spatial memory task provides evidence of stimuli differentiation.
****

ETNet: Error Transition Network for Arbitrary Style Transfer
Chunjin Song, Zhijie Wu, Yang Zhou, Minglun Gong, Hui Huang
Numerous valuable efforts have been devoted to achieving arbitrary style transfer since the seminal work of Gatys et al. However, existing state-of-the-art approaches often generate insufficiently stylized results under challenging cases. We believe a fundamental reason is that these approaches try to generate the stylized result in a single shot and hence fail to fully satisfy the constraints on semantic structures in the content images and style patterns in the style images. Inspired by the works on error-correction, instead, we propose a self-correcting model to predict what is wrong with the current stylization and refine it accordingly in an iterative manner. For each refinement, we transit the error features across both the spatial and scale domain and invert the processed features into a residual image, with a network we call Error Transition Network (ETNet). The proposed model improves over the state-of-the-art methods with better semantic structures and more adaptive style pattern details. Various qualitative and quantitative experiments show that the key concept of both progressive strategy and error-correction leads to better results. Code and models are available at https://github.com/zhijieW94/ETNet.
****

Cross-lingual Language Model Pretraining
Alexis CONNEAU, Guillaume Lample
Recent studies have demonstrated the efficiency of generative pretraining for English natural language understanding. In this work, we extend this approach to multiple languages and show the effectiveness of cross-lingual pretraining. We propose two methods to learn cross-lingual language models (XLMs): one unsupervised that only relies on monolingual data, and one supervised that leverages parallel data with a new cross-lingual language model objective. We obtain state-of-the-art results on cross-lingual classification, unsupervised and supervised machine translation. On XNLI, our approach pushes the state of the art by an absolute gain of 4.9% accuracy. On unsupervised machine translation, we obtain 34.3 BLEU on WMT'16 German-English, improving the previous state of the art by more than 9 BLEU. On supervised machine translation, we obtain a new state of the art of 38.5 BLEU on WMT'16 Romanian-English, outperforming the previous best approach by more than 4 BLEU. Our code and pretrained models will be made publicly available.
****

Icebreaker: Element-wise Efficient Information Acquisition with a Bayesian Deep Latent Gaussian Model
Wenbo Gong, Sebastian Tschiatschek, Sebastian Nowozin, Richard E. Turner, José Miguel Hernández-Lobato, Cheng Zhang
In this paper, we address the ice-start problem, i.e., the challenge of deploying machine learning models when only a little or no training data is initially available, and acquiring each feature element of data is associated with costs. This setting is representative of the real-world machine learning applications. For instance, in the health care domain, obtaining every single measurement comes with a cost. We propose Icebreaker, a principled framework for elementwise training data acquisition. Icebreaker introduces a full Bayesian Deep Latent Gaussian Model (BELGAM) with a novel inference method, which combines recent advances in amortized inference and stochastic gradient MCMC to enable fast and accurate po

sterior inference. By utilizing BELGAM’s ability to fully quantify model uncertainty, we also propose two information acquisition functions for imputation and active prediction problems. We demonstrate that BELGAM performs significantly better than previous variational autoencoder (VAE) based models, when the data set size is small, using both machine learning benchmarks and real world recommender systems and health-care applications. Moreover, Icebreaker not only demonstrates improved performance compared to baselines, but it is also capable of achieving better test performance with less training data available.
****

Efficient and Thrifty Voting by Any Means Necessary
Debmalya Mandal, Ariel D. Procaccia, Nisarg Shah, David Woodruff
We take an unorthodox view of voting by expanding the design space to include both the elicitation rule, whereby voters map their (cardinal) preferences to votes, and the aggregation rule, which transforms the reported votes into collective decisions. Intuitively, there is a tradeoff between the communication requirements of the elicitation rule (i.e., the number of bits of information that voters need to provide about their preferences) and the efficiency of the outcome of the aggregation rule, which we measure through distortion (i.e., how well the utilitarian social welfare of the outcome approximates the maximum social welfare in the worst case). Our results chart the Pareto frontier of the communication-distortion tradeoff.
****

Post training 4-bit quantization of convolutional networks for rapid-deployment
Ron Banner, Yury Nahshan, Daniel Soudry
Convolutional neural networks require significant memory bandwidth and storage for intermediate computations, apart from substantial computing resources. Neural network quantization has significant benefits in reducing the amount of intermediate results, but it often requires the full datasets and time-consuming fine tuning to recover the accuracy lost after quantization. This paper introduces the first practical 4-bit post training quantization approach: it does not involve training the quantized model (fine-tuning), nor it requires the availability of the full dataset. We target the quantization of both activations and weights and suggest three complementary methods for minimizing quantization error at the tensor level, two of whom obtain a closed-form analytical solution. Combining these methods, our approach achieves accuracy that is just a few percents less the state-of-the-art baseline across a wide range of convolutional models. The source code to replicate all experiments is available on GitHub: \url{https://github.com/submission2019/cnn-quantization}.
****

Implicit Regularization in Deep Matrix Factorization
Sanjeev Arora, Nadav Cohen, Wei Hu, Yuping Luo
Efforts to understand the generalization mystery in deep learning have led to the belief that gradient-based optimization induces a form of implicit regularization, a bias towards models of low "complexity." We study the implicit regularization of gradient descent over deep linear neural networks for matrix completion and sensing, a model referred to as deep matrix factorization. Our first finding, supported by theory and experiments, is that adding depth to a matrix factorization enhances an implicit tendency towards low-rank solutions, oftentimes leading to more accurate recovery. Secondly, we present theoretical and empirical arguments questioning a nascent view by which implicit regularization in matrix factorization can be captured using simple mathematical norms. Our results point to the possibility that the language of standard regularizers may not be rich enough to fully encompass the implicit regularization brought forth by gradient-based optimization.
****

Crowdsourcing via Pairwise Co-occurrences: Identifiability and Algorithms
Shahana Ibrahim, Xiao Fu, Nikolaos Kargas, Kejun Huang
The data deluge comes with high demands for data labeling. Crowdsourcing (or, more generally, ensemble learning) techniques aim to produce accurate labels via integrating noisy, non-expert labeling from annotators. The classic Dawid-Skene e

stimator and its accompanying expectation maximization (EM) algorithm have been widely used, but the theoretical properties are not fully understood. Tensor methods were proposed to guarantee identification of the Dawid-Skene model, but the sample complexity is a hurdle for applying such approaches---since the tensor methods hinge on the availability of third-order statistics that are hard to reliably estimate given limited data. In this paper, we propose a framework using pairwise co-occurrences of the annotator responses, which naturally admits lower sample complexity. We show that the approach can identify the Dawid-Skene model under realistic conditions. We propose an algebraic algorithm reminiscent of convex geometry-based structured matrix factorization to solve the model identification problem efficiently, and an identifiability-enhanced algorithm for handling more challenging and critical scenarios. Experiments show that the proposed algorithms outperform the state-of-art algorithms under a variety of scenarios.
****
Learning low-dimensional state embeddings and metastable clusters from time series data
Yifan Sun, Yaqi Duan, Hao Gong, Mengdi Wang
This paper studies how to find compact state embeddings from high-dimensional Markov state trajectories, where the transition kernel has a small intrinsic rank. In the spirit of diffusion map, we propose an efficient method for learning a low-dimensional state embedding and capturing the process's dynamics. This idea also leads to a kernel reshaping method for more accurate nonparametric estimation of the transition function. State embedding can be used to cluster states into metastable sets, thereby identifying the slow dynamics. Sharp statistical error bounds and misclassification rate are proved. Experiment on a simulated dynamical system shows that the state clustering method indeed reveals metastable structures. We also experiment with time series generated by layers of a Deep-Q-Network when playing an Atari game. The embedding method identifies game states to be similar if they share similar future events, even though their raw data are far different.
****
Necessary and Sufficient Geometries for Gradient Methods
Daniel Levy, 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.
****
Limitations of Lazy Training of Two-layers Neural Network
Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, Andrea Montanari
We study the supervised learning problem under either of the following two models:
(1) Feature vectors xi are d-dimensional Gaussian and responses are yi = f*(xi) for f* an unknown quadratic function;
(2) Feature vectors xi are distributed as a mixture of two d-dimensional centered Gaussians, and y_i's are the corresponding class labels.
We use two-layers neural networks with quadratic activations, and compare three different learning regimes: the random features (RF) regime in which we only train the second-layer weights; the neural tangent (NT) regime in which we train a linearization of the neural network around its initialization; the fully trained neural network (NN) regime in which we train all the weights in the network. We prove that, even for the simple quadratic model of point (1), there is a potentially unbounded gap between the prediction risk achieved in these three training regimes, when the number of neurons is smaller than the ambient dimension. When the number of neurons is larger than the number of dimensions, the problem is significantly easier and both NT and NN learning achieve zero risk.
****
Learning Auctions with Robust Incentive Guarantees
Jacob D. Abernethy, Rachel Cummings, Bhuvesh Kumar, Sam Taggart, Jamie H. Morgenstern

We study the problem of learning Bayesian-optimal revenue-maximizing auctions. The classical approach to maximizing revenue requires a known prior distribution on the demand of the bidders, although recent work has shown how to replace the knowledge of a prior distribution with a polynomial sample. However, in an online setting, when buyers can participate in multiple rounds, standard learning techniques are susceptible to \emph{strategic overfitting}: bidders can improve their long-term wellbeing by manipulating the trajectory of the learning algorithm in earlier rounds. For example, they may be able to strategically adjust their behavior in earlier rounds to achieve lower, more favorable future prices. Such non-truthful behavior can hinder learning and harm revenue. In this paper, we combine tools from differential privacy, mechanism design, and sample complexity to give a repeated auction that (1) learns bidder demand from past data, (2) is approximately revenue-optimal, and (3) strategically robust, as it incentivizes bidders to behave truthfully.
****
Local SGD with Periodic Averaging: Tighter Analysis and Adaptive Synchronization
Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, Viveck Cadambe
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.
****
Scalable Bayesian inference of dendritic voltage via spatiotemporal recurrent state space models
Ruoxi Sun, Scott Linderman, Ian Kinsella, Liam Paninski
Recent advances in optical voltage sensors have brought us closer to a critical goal in cellular neuroscience: imaging the full spatiotemporal voltage on a dendritic tree. However, current sensors and imaging approaches still face significant limitations in SNR and sampling frequency; therefore statistical denoising and interpolation methods remain critical for understanding single-trial spatiotemporal dendritic voltage dynamics. Previous denoising approaches were either based on an inadequate linear voltage model or scaled poorly to large trees. Here we introduce a scalable fully Bayesian approach. We develop a generative nonlinear model that requires few parameters per compartment of the cell but is nonetheless flexible enough to sample realistic spatiotemporal data. The model captures different dynamics in each compartment and leverages biophysical knowledge to constrain intra- and inter-compartmental dynamics. We obtain a full posterior distribution over spatiotemporal voltage via an augmented Gibbs sampling algorithm. The nonlinear smoother model outperforms previously developed linear methods, and scales to much larger systems than previous methods based on sequential Monte Carlo approaches.
****
Constrained Reinforcement Learning Has Zero Duality Gap
Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, Alejandro Ribeiro
Autonomous agents must often deal with conflicting requirements, such as completing tasks using the least amount of time/energy, learning multiple tasks, or dealing with multiple opponents. In the context of reinforcement learning~(RL), these problems are addressed by (i)~designing a reward function that simultaneously describes all requirements or (ii)~combining modular value functions that encode them individually. Though effective, these methods have critical downsides. Designing good reward functions that balance different objectives is challenging, especially as the number of objectives grows. Moreover, implicit interference between goals may lead to performance plateaus as they compete for resources, particularly when training on-policy. Similarly, selecting parameters to combine value functions is at least as hard as designing an all-encompassing reward, given that the effect of their values on the overall policy is not straightforward. The later is generally addressed by formulating the conflicting requirements as a constrained RL problem and solved using Primal-Dual methods. These algorithms are in general not guaranteed to converge to the optimal solution since the proble

m is not convex. This work provides theoretical support to these approaches by e
stablishing that despite its non-convexity, this problem has zero duality gap, i
.e., it can be solved exactly in the dual domain, where it becomes convex. Final
ly, we show this result basically holds if the policy is described by a good par
ametrization~(e.g., neural networks) and we connect this result with primal-dual
 algorithms present in the literature and we establish the convergence to the op
timal solution.
****
A Meta-MDP Approach to Exploration for Lifelong Reinforcement Learning
Francisco Garcia, Philip S. Thomas
In this paper we consider the problem of how a reinforcement learning agent that
 is tasked with solving a sequence of reinforcement learning problems (a sequenc
e of Markov decision processes) can use knowledge acquired early in its lifetime
 to improve its ability to solve new problems. We argue that previous experience
 with similar problems can provide an agent with information about how it should
 explore when facing a new but related problem. We show that the search for an o
ptimal exploration strategy can be formulated as a reinforcement learning proble
m itself and demonstrate that such strategy can leverage patterns found in the s
tructure of related problems.
We conclude with experiments that show the benefits of optimizing an exploration
 strategy using our proposed framework.
****
Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction
Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, Sergey Levine
Off-policy reinforcement learning aims to leverage experience collected from pri
or policies for sample-efficient learning. However, in practice, commonly used o
ff-policy approximate dynamic programming methods based on Q-learning and actor-
critic methods are highly sensitive to the data distribution, and can make only
limited progress without collecting additional on-policy data. As a step towards
 more robust off-policy algorithms, we study the setting where the off-policy ex
perience is fixed and there is no further interaction with the environment. We i
dentify \emph{bootstrapping error} as a key source of instability in current met
hods. Bootstrapping error is due to bootstrapping from actions that lie outside
of the training data distribution, and it accumulates via the Bellman backup ope
rator. We theoretically analyze bootstrapping error, and demonstrate how careful
ly constraining action selection in the backup can mitigate it. Based on our ana
lysis, we propose a practical algorithm, bootstrapping error accumulation reduct
ion (BEAR). We demonstrate that BEAR is able to learn robustly from different of
f-policy distributions, including random data and suboptimal demonstrations, on
a range of continuous control tasks.
****
Learning by Abstraction: The Neural State Machine
Drew Hudson, Christopher D. Manning
We introduce the Neural State Machine, seeking to bridge the gap between the neu
ral and symbolic views of AI and integrate their complementary strengths for the
 task of visual reasoning. Given an image, we first predict a probabilistic grap
h that represents its underlying semantics and serves as a structured world mode
l. Then, we perform sequential reasoning over the graph, iteratively traversing
its nodes to answer a given question or draw a new inference. In contrast to mos
t neural architectures that are designed to closely interact with the raw sensor
y data, our model operates instead in an abstract latent space, by transforming
both the visual and linguistic modalities into semantic concept-based representa
tions, thereby achieving enhanced transparency and modularity. We evaluate our m
odel on VQA-CP and GQA, two recent VQA datasets that involve compositionality, m
ulti-step inference and diverse reasoning skills, achieving state-of-the-art res
ults in both cases. We provide further experiments that illustrate the model's s
trong generalization capacity across multiple dimensions, including novel compos
itions of concepts, changes in the answer distribution, and unseen linguistic st
ructures, demonstrating the qualities and efficacy of our approach.
****

Unified Language Model Pre-training for Natural Language Understanding and Generation
Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, Hsiao-Wuen Hon
This paper presents a new Unified pre-trained Language Model (UniLM) that can be fine-tuned for both natural language understanding and generation tasks. The model is pre-trained using three types of language modeling tasks: unidirectional, bidirectional, and sequence-to-sequence prediction. The unified modeling is achieved by employing a shared Transformer network and utilizing specific self-attention masks to control what context the prediction conditions on. UniLM compares favorably with BERT on the GLUE benchmark, and the SQuAD 2.0 and CoQA question answering tasks. Moreover, UniLM achieves new state-of-the-art results on five natural language generation datasets, including improving the CNN/DailyMail abstractive summarization ROUGE-L to 40.51 (2.04 absolute improvement), the Gigaword abstractive summarization ROUGE-L to 35.75 (0.86 absolute improvement), the CoQA generative question answering F1 score to 82.5 (37.1 absolute improvement), the SQuAD question generation BLEU-4 to 22.12 (3.75 absolute improvement), and the DSTC7 document-grounded dialog response generation NIST-4 to 2.67 (human performance is 2.65). The code and pre-trained models are available at https://github.com/microsoft/unilm.
****
Adaptive GNN for Image Analysis and Editing
Lingyu Liang, LianWen Jin, Yong Xu
Graph neural network (GNN) has powerful representation ability, but optimal configurations of GNN are non-trivial to obtain due to diversity of graph structure and cascaded nonlinearities. This paper aims to understand some properties of GNN from a computer vision (CV) perspective. In mathematical analysis, we propose an adaptive GNN model by recursive definition, and derive its relation with two basic operations in CV: filtering and propagation operations. The proposed GNN model is formulated as a label propagation system with guided map, graph Laplacian and node weight. It reveals that 1) the guided map and node weight determine whether a GNN leads to filtering or propagation diffusion, and 2) the kernel of graph Laplacian controls diffusion pattern. In practical verification, we design a new regularization structure with guided feature to produce GNN-based filtering and propagation diffusion to tackle the ill-posed inverse problems of quotient image analysis (QIA), which recovers the reflectance ratio as a signature for image analysis or adjustment. A flexible QIA-GNN framework is constructed to achieve various image-based editing tasks, like face illumination synthesis and low-light image enhancement. Experiments show the effectiveness of the QIA-GNN, and provide new insights of GNN for image analysis and editing.
****
Metric Learning for Adversarial Robustness
Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, Baishakhi Ray
Deep networks are well-known to be fragile to adversarial attacks. We conduct an empirical analysis of deep representations under the state-of-the-art attack method called PGD, and find that the attack causes the internal representation to shift closer to the ``false'' class. Motivated by this observation, we propose to regularize the representation space under attack with metric learning to produce more robust classifiers. By carefully sampling examples for metric learning, our learned representation not only increases robustness, but also detects previously unseen adversarial samples. Quantitative experiments show improvement of robustness accuracy by up to 4% and detection efficiency by up to 6% according to Area Under Curve score over prior work. The code of our work is available at https://github.com/columbia/MetricLearningAdversarial_Robustness.
****
Fine-grained Optimization of Deep Neural Networks
Mete Ozay
In recent studies, several asymptotic upper bounds on generalization errors on deep neural networks (DNNs) are theoretically derived. These bounds are functions of several norms of weights of the DNNs, such as the Frobenius and spectral nor

ms, and they are computed for weights grouped according to either input and outp ut channels of the DNNs. In this work, we conjecture that if we can impose multi ple constraints on weights of DNNs to upper bound the norms of the weights, and train the DNNs with these weights, then we can attain empirical generalization e rrors closer to the derived theoretical bounds, and improve accuracy of the DNNs .
****
Learning to Control Self-Assembling Morphologies: A Study of Generalization via Modularity
Deepak Pathak, Christopher Lu, Trevor Darrell, Phillip Isola, Alexei A. Efros
Contemporary sensorimotor learning approaches typically start with an existing c omplex agent (e.g., a robotic arm), which they learn to control. In contrast, th is paper investigates a modular co-evolution strategy: a collection of primitive agents learns to dynamically self-assemble into composite bodies while also lea rning to coordinate their behavior to control these bodies. Each primitive agent consists of a limb with a motor attached at one end. Limbs may choose to link u p to form collectives. When a limb initiates a link-up action and there is anoth er limb nearby, the latter is magnetically connected to the 'parent' limb's moto r. This forms a new single agent, which may further link with other agents. In t his way, complex morphologies can emerge, controlled by a policy whose architect ure is in explicit correspondence with the morphology. We evaluate the performan ce of these dynamic and modular agents in simulated environments. We demonstrate better generalization to test-time changes both in the environment, as well as in the structure of the agent, compared to static and monolithic baselines. Proj ect videos and source code are provided in the supplementary material.
****
An adaptive Mirror-Prox method for variational inequalities with singular operat ors
Kimon Antonakopoulos, Veronica Belmega, Panayotis Mertikopoulos
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.
****
Alleviating Label Switching with Optimal Transport
Pierre Monteiller, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, Justin M . Solomon, Mikhail Yurochkin
Label switching is a phenomenon arising in mixture model posterior inference tha t prevents one from meaningfully assessing posterior statistics using standard M onte Carlo procedures. This issue arises due to invariance of the posterior unde r actions of a group; for example, permuting the ordering of mixture components has no effect on the likelihood. We propose a resolution to label switching that leverages machinery from optimal transport. Our algorithm efficiently computes posterior statistics in the quotient space of the symmetry group. We give condi tions under which there is a meaningful solution to label switching and demonstr ate advantages over alternative approaches on simulated and real data.
****
Fisher Efficient Inference of Intractable Models
Song Liu, Takafumi Kanamori, Wittawat Jitkrittum, Yu Chen
Maximum Likelihood Estimators (MLE) has many good properties. For example, the a symptotic variance of MLE solution attains equality of the asymptotic Cram{\'e}r -Rao lower bound (efficiency bound), which is the minimum possible variance for an unbiased estimator. However, obtaining such MLE solution requires calculating the likelihood function which may not be tractable due to the normalization ter m of the density model. In this paper, we derive a Discriminative Likelihood Est imator (DLE) from the Kullback-Leibler divergence minimization criterion impleme nted via density ratio estimation and a Stein operator. We study the problem of model inference using DLE. We prove its consistency and show that the asymptotic variance of its solution can attain the equality of the efficiency bound under mild regularity conditions. We also propose a dual formulation of DLE which can

be easily optimized. Numerical studies validate our asymptotic theorems and we give an example where DLE successfully estimates an intractable model constructed using a pre-trained deep neural network.
****
Stochastic Gradient Hamiltonian Monte Carlo Methods with Recursive Variance Reduction
Difan Zou, Pan Xu, Quanquan Gu
Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) algorithms have received increasing attention in both theory and practice. In this paper, we propose a Stochastic Recursive Variance-Reduced gradient HMC (SRVR-HMC) algorithm. It makes use of a semi-stochastic gradient estimator that recursively accumulates the gradient information to reduce the variance of the stochastic gradient. We provide a convergence analysis of SRVR-HMC for sampling from a class of non-log-concave distributions and show that SRVR-HMC converges faster than all existing HMC-type algorithms based on underdamped Langevin dynamics. Thorough experiments on synthetic and real-world datasets validate our theory and demonstrate the superiority of SRVR-HMC.
****
Online Learning via the Differential Privacy Lens
Jacob D. Abernethy, Young Hun Jung, Chansoo Lee, Audra McMillan, Ambuj Tewari
In this paper, we use differential privacy as a lens to examine online learning in both full and partial information settings. The differential privacy framework is, at heart, less about privacy and more about algorithmic stability, and thus has found application in domains well beyond those where information security is central. Here we develop an algorithmic property called one-step differential stability which facilitates a more refined regret analysis for online learning methods. We show that tools from the differential privacy literature can yield regret bounds for many interesting online learning problems including online convex optimization and online linear optimization. Our stability notion is particularly well-suited for deriving first-order regret bounds for follow-the-perturbed-leader algorithms, something that all previous analyses have struggled to achieve. We also generalize the standard max-divergence to obtain a broader class called Tsallis max-divergences. These define stronger notions of stability that are useful in deriving bounds in partial information settings such as multi-armed bandits and bandits with experts.
****
Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions
Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, Elias Bareinboim
The challenge of learning the causal structure underlying a certain phenomenon is undertaken by connecting the set of conditional independences (CIs) readable from the observational data, on the one side, with the set of corresponding constraints implied over the graphical structure, on the other, which are tied through a graphical criterion known as d-separation (Pearl, 1988). In this paper, we investigate the more general scenario where multiple observational and experimental distributions are available. We start with the simple observation that the invariances given by CIs/d-separation are just one special type of a broader set of constraints, which follow from the careful comparison of the different distributions available. Remarkably, these new constraints are intrinsically connected with do-calculus (Pearl, 1995) in the context of soft-interventions. We introduce a novel notion of interventional equivalence class of causal graphs with latent variables based on these invariances, which associates each graphical structure with a set of interventional distributions that respect the do-calculus rules. Given a collection of distributions, two causal graphs are called interventionally equivalent if they are associated with the same family of interventional distributions, where the elements of the family are indistinguishable using the invariances obtained from a direct application of the calculus rules. We introduce a graphical representation that can be used to determine if two causal graphs are interventionally equivalent. We provide a formal graphical characterization of this equivalence. Finally, we extend the FCI algorithm, which was originally d

esigned to operate based on CIs, to combine observational and interventional dat
asets, including new orientation rules particular to this setting.
****
Domes to Drones: Self-Supervised Active Triangulation for 3D Human Pose Reconstr
uction
Aleksis Pirinen, Erik Gärtner, Cristian Sminchisescu
Existing state-of-the-art estimation systems can detect 2d poses of multiple peo
ple in images quite reliably. In contrast, 3d pose estimation from a single imag
e is ill-posed due to occlusion and depth ambiguities. Assuming access to multip
le cameras, or given an active system able to position itself to observe the sce
ne from multiple viewpoints, reconstructing 3d pose from 2d measurements becomes
 well-posed within the framework of standard multi-view geometry. Less clear is
what is an informative set of viewpoints for accurate 3d reconstruction, particu
larly in complex scenes, where people are occluded by others or by scene objects
. In order to address the view selection problem in a principled way, we here in
troduce ACTOR, an active triangulation agent for 3d human pose reconstruction. O
ur fully trainable agent consists of a 2d pose estimation network (any of which
would work) and a deep reinforcement learning-based policy for camera viewpoint
selection. The policy predicts observation viewpoints, the number of which varie
s adaptively depending on scene content, and the associated images are fed to an
 underlying pose estimator. Importantly, training the policy requires no annotat
ions - given a 2d pose estimator, ACTOR is trained in a self-supervised manner.
In extensive evaluations on complex multi-people scenes filmed in a Panoptic dom
e, under multiple viewpoints, we compare our active triangulation agent to stron
g multi-view baselines, and show that ACTOR produces significantly more accurate
 3d pose reconstructions. We also provide a proof-of-concept experiment indicati
ng the potential of connecting our view selection policy to a physical drone obs
erver.
****
SIC-MMAB: Synchronisation Involves Communication in Multiplayer Multi-Armed Band
its
Etienne Boursier, Vianney Perchet
Motivated by cognitive radio networks, we consider the stochastic multiplayer mu
lti-armed bandit problem, where several players pull arms simultaneously and col
lisions occur if one of them is pulled by several players at the same stage. We
 present a decentralized algorithm that achieves the same performance as a centr
alized one, contradicting the existing lower bounds for that problem. This is p
ossible by ``hacking'' the standard model by constructing a communication protoc
ol between players that deliberately enforces collisions, allowing them to share
 their information at a negligible cost.
This motivates the introduction of a more appropriate dynamic setting without se
nsing, where similar communication protocols are no longer possible. However, we
 show that the logarithmic growth of the regret is still achievable for this mod
el with a new algorithm.
****
A Step Toward Quantifying Independently Reproducible Machine Learning Research
Edward Raff
What makes a paper independently reproducible? Debates on reproducibility center
 around intuition or assumptions but lack empirical results. Our field focuses o
n releasing code, which is important, but is not sufficient for determining repr
oducibility. We take the first step toward a quantifiable answer by manually att
empting to implement 255 papers published from 1984 until 2017, recording featur
es of each paper, and performing statistical analysis of the results. For each p
aper, we did not look at the authors code, if released, in order to prevent bias
 toward discrepancies between code and paper.
****
Latent distance estimation for random geometric graphs
Ernesto Araya Valdivia, De Castro Yohann
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 Adversarial Semantics-Consistent Network for Generalized Zero-Shot Learning
Jian Ni, Shanghang Zhang, Haiyong Xie
Generalized zero-shot learning (GZSL) is a challenging class of vision and knowledge transfer problems in which both seen and unseen classes appear during testing. Existing GZSL approaches either suffer from semantic loss and discard discriminative information at the embedding stage, or cannot guarantee the visual-semantic interactions. To address these limitations, we propose a Dual Adversarial Semantics-Consistent Network (referred to as DASCN), which learns both primal and dual Generative Adversarial Networks (GANs) in a unified framework for GZSL. In DASCN, the primal GAN learns to synthesize inter-class discriminative and semantics-preserving visual features from both the semantic representations of seen/unseen classes and the ones reconstructed by the dual GAN. The dual GAN enforces the synthetic visual features to represent prior semantic knowledge well via semantics-consistent adversarial learning. To the best of our knowledge, this is the first work that employs a novel dual-GAN mechanism for GZSL. Extensive experiments show that our approach achieves significant improvements over the state-of-the-art approaches.
****

Manipulating a Learning Defender and Ways to Counteract
Jiarui Gan, Qingyu Guo, Long Tran-Thanh, Bo An, Michael Wooldridge
In Stackelberg security games when information about the attacker's payoffs is uncertain, algorithms have been proposed to learn the optimal defender commitment by interacting with the attacker and observing their best responses. In this paper, we show that, however, these algorithms can be easily manipulated if the attacker responds untruthfully. As a key finding, attacker manipulation normally leads to the defender learning a maximin strategy, which effectively renders the learning attempt meaningless as to compute a maximin strategy requires no additional information about the other player at all. We then apply a game-theoretic framework at a higher level to counteract such manipulation, in which the defender commits to a policy that specifies her strategy commitment according to the learned information. We provide a polynomial-time algorithm to compute the optimal such policy, and in addition, a heuristic approach that applies even when the attacker's payoff space is infinite or completely unknown. Empirical evaluation shows that our approaches can improve the defender's utility significantly as compared to the situation when attacker manipulation is ignored.
****

Privacy Amplification by Mixing and Diffusion Mechanisms
Borja Balle, Gilles Barthe, Marco Gaboardi, Joseph Geumlek
A fundamental result in differential privacy states that the privacy guarantees of a mechanism are preserved by any post-processing of its output. In this paper we investigate under what conditions stochastic post-processing can amplify the privacy of a mechanism. By interpreting post-processing as the application of a Markov operator, we first give a series of amplification results in terms of uniform mixing properties of the Markov process defined by said operator. Next we provide amplification bounds in terms of coupling arguments which can be applied in cases where uniform mixing is not available. Finally, we introduce a new family of mechanisms based on diffusion processes which are closed under post-processing, and analyze their privacy via a novel heat flow argument. On the applied side, we generalize the analysis of "privacy amplification by iteration" in Noisy SGD and show it admits an exponential improvement in the strongly convex case, and study a mechanism based on the Ornstein–Uhlenbeck diffusion process which contains the Gaussian mechanism with optimal post-processing on bounded inputs as a special case.
****

Ultra Fast Medoid Identification via Correlated Sequential Halving
Tavor Baharav, David Tse
The medoid of a set of n points is the point in the set that minimizes the sum o

f distances to other points. It can be determined exactly in O(n^2) time by computing the distances between all pairs of points. Previous works show that one can significantly reduce the number of distance computations needed by adaptively querying distances. The resulting randomized algorithm is obtained by a direct conversion of the computation problem to a multi-armed bandit statistical inference problem. In this work, we show that we can better exploit the structure of the underlying computation problem by modifying the traditional bandit sampling strategy and using it in conjunction with a suitably chosen multi-armed bandit algorithm. Four to five orders of magnitude gains over exact computation are obtained on real data, in terms of both number of distance computations needed and wall clock time. Theoretical results are obtained to quantify such gains in terms of data parameters. Our code is publicly available online at https://github.com/TavorB/Correlated-Sequential-Halving.
****

On the Inductive Bias of Neural Tangent Kernels

Alberto Bietti, Julien Mairal

State-of-the-art neural networks are heavily over-parameterized, making the optimization algorithm a crucial ingredient for learning predictive models with good generalization properties. A recent line of work has shown that in a certain over-parameterized regime, the learning dynamics of gradient descent are governed by a certain kernel obtained at initialization, called the neural tangent kernel. We study the inductive bias of learning in such a regime by analyzing this kernel and the corresponding function space (RKHS). In particular, we study smoothness, approximation, and stability properties of functions with finite norm, including stability to image deformations in the case of convolutional networks, and compare to other known kernels for similar architectures.
****

Surround Modulation: A Bio-inspired Connectivity Structure for Convolutional Neural Networks

Hosein Hasani, Mahdieh Soleymani, Hamid Aghajan

Numerous neurophysiological studies have revealed that a large number of the primary visual cortex neurons operate in a regime called surround modulation. Surround modulation has a substantial effect on various perceptual tasks, and it also plays a crucial role in the efficient neural coding of the visual cortex. Inspired by the notion of surround modulation, we designed new excitatory-inhibitory connections between a unit and its surrounding units in the convolutional neural network (CNN) to achieve a more biologically plausible network. Our experiments show that this simple mechanism can considerably improve both the performance and training speed of traditional CNNs in visual tasks. We further explore additional outcomes of the proposed structure. We first evaluate the model under several visual challenges, such as the presence of clutter or change in lighting conditions and show its superior generalization capability in handling these challenging situations. We then study possible changes in the statistics of neural activities such as sparsity and decorrelation and provide further insight into the underlying efficiencies of surround modulation. Experimental results show that importing surround modulation into the convolutional layers ensues various effects analogous to those derived by surround modulation in the visual cortex.
****

Rethinking Kernel Methods for Node Representation Learning on Graphs

Yu Tian, Long Zhao, Xi Peng, Dimitris Metaxas

Graph kernels are kernel methods measuring graph similarity and serve as a standard tool for graph classification. However, the use of kernel methods for node classification, which is a related problem to graph representation learning, is still ill-posed and the state-of-the-art methods are heavily based on heuristics. Here, we present a novel theoretical kernel-based framework for node classification that can bridge the gap between these two representation learning problems on graphs. Our approach is motivated by graph kernel methodology but extended to learn the node representations capturing the structural information in a graph. We theoretically show that our formulation is as powerful as any positive semidefinite kernels. To efficiently learn the kernel, we propose a novel mechanism f

or node feature aggregation and a data-driven similarity metric employed during the training phase. More importantly, our framework is flexible and complementary to other graph-based deep learning models, e.g., Graph Convolutional Networks (GCNs). We empirically evaluate our approach on a number of standard node classification benchmarks, and demonstrate that our model sets the new state of the art.
****
A Necessary and Sufficient Stability Notion for Adaptive Generalization
Moshe Shenfeld, Katrina Ligett
We introduce a new notion of the stability of computations, which holds under post-processing and adaptive composition. We show that the notion is both necessary and sufficient to ensure generalization in the face of adaptivity, for any computations that respond to bounded-sensitivity linear queries while providing accuracy with respect to the data sample set. The stability notion is based on quantifying the effect of observing a computation's outputs on the posterior over the data sample elements. We show a separation between this stability notion and previously studied notion and observe that all differentially private algorithms also satisfy this notion.
****
Implicit Regularization of Accelerated Methods in Hilbert Spaces
Nicolò Pagliana, Lorenzo Rosasco
We study learning properties of accelerated gradient descent methods for linear least-squares in Hilbert spaces. We analyze the implicit regularization properties of Nesterov acceleration and a variant of heavy-ball in terms of corresponding learning error bounds. Our results show that acceleration can provides faster bias decay than gradient descent, but also suffers of a more unstable behavior. As a result acceleration cannot be in general expected to improve learning accuracy with respect to gradient descent, but rather to achieve the same accuracy with reduced computations. Our theoretical results are validated by numerical simulations. Our analysis is based on studying suitable polynomials induced by the accelerated dynamics and combining spectral techniques with concentration inequalities.
****
Input Similarity from the Neural Network Perspective
Guillaume Charpiat, Nicolas Girard, Loris Felardos, Yuliya Tarabalka
Given a trained neural network, we aim at understanding how similar it considers any two samples. For this, we express a proper definition of similarity from the neural network perspective (i.e. we quantify how undissociable two inputs A and B are), by taking a machine learning viewpoint: how much a parameter variation designed to change the output for A would impact the output for B as well?
****
Transfer Learning via Minimizing the Performance Gap Between Domains
Boyu Wang, Jorge Mendez, Mingbo Cai, Eric Eaton
We propose a new principle for transfer learning, based on a straightforward intuition: if two domains are similar to each other, the model trained on one domain should also perform well on the other domain, and vice versa. To formalize this intuition, we define the performance gap as a measure of the discrepancy between the source and target domains. We derive generalization bounds for the instance weighting approach to transfer learning, showing that the performance gap can be viewed as an algorithm-dependent regularizer, which controls the model complexity. Our theoretical analysis provides new insight into transfer learning and motivates a set of general, principled rules for designing new instance weighting schemes for transfer learning. These rules lead to gapBoost, a novel and principled boosting approach for transfer learning. Our experimental evaluation on benchmark data sets shows that gapBoost significantly outperforms previous boosting-based transfer learning algorithms.
****
Catastrophic Forgetting Meets Negative Transfer: Batch Spectral Shrinkage for Safe Transfer Learning
Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, Jianmin Wang

Before sufficient training data is available, fine-tuning neural networks pre-trained on large-scale datasets substantially outperforms training from random initialization. However, fine-tuning methods suffer from two dilemmas, catastrophic forgetting and negative transfer. While several methods with explicit attempts to overcome catastrophic forgetting have been proposed, negative transfer is rarely delved into. In this paper, we launch an in-depth empirical investigation into negative transfer in fine-tuning and find that, for the weight parameters and feature representations, transferability of their spectral components is diverse. For safe transfer learning, we present Batch Spectral Shrinkage (BSS), a novel regularization approach to penalizing smaller singular values so that untransferable spectral components are suppressed. BSS is orthogonal to existing fine-tuning methods and is readily pluggable to them. Experimental results show that BSS can significantly enhance the performance of representative methods, especially with limited training data.
****
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks
Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee
We present ViLBERT (short for Vision-and-Language BERT), a model for learning task-agnostic joint representations of image content and natural language. We extend the popular BERT architecture to a multi-modal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers. We pretrain our model through two proxy tasks on the large, automatically collected Conceptual Captions dataset and then transfer it to multiple established vision-and-language tasks -- visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval -- by making only minor additions to the base architecture. We observe significant improvements across tasks compared to existing task-specific models -- achieving state-of-the-art on all four tasks. Our work represents a shift away from learning groundings between vision and language only as part of task training and towards treating visual grounding as a pretrainable and transferable capability.
****
Efficiently Learning Fourier Sparse Set Functions
Andisheh Amrollahi, Amir Zandieh, Michael Kapralov, Andreas Krause
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.
****
Search-Guided, Lightly-Supervised Training of Structured Prediction Energy Networks
Amirmohammad Rooshenas, Dongxu Zhang, Gopal Sharma, Andrew McCallum
In structured output prediction tasks, labeling ground-truth training output is often expensive. However, for many tasks, even when the true output is unknown, we can evaluate predictions using a scalar reward function, which may be easily assembled from human knowledge or non-differentiable pipelines. But searching through the entire output space to find the best output with respect to this reward function is typically intractable. In this paper, we instead use efficient truncated randomized search in this reward function to train structured prediction energy networks (SPENs), which provide efficient test-time inference using gradient-based search on a smooth, learned representation of the score landscape, and have previously yielded state-of-the-art results in structured prediction. In particular, this truncated randomized search in the reward function yields previously unknown local improvements, providing effective supervision to SPENs, avoiding their traditional need for labeled training data.
****
Planning with Goal-Conditioned Policies
Soroush Nasiriany, Vitchyr Pong, Steven Lin, Sergey Levine
Planning methods can solve temporally extended sequential decision making proble

ms by composing simple behaviors. However, planning requires suitable abstractions for the states and transitions, which typically need to be designed by hand. In contrast, reinforcement learning (RL) can acquire behaviors from low-level inputs directly, but struggles with temporally extended tasks. Can we utilize reinforcement learning to automatically form the abstractions needed for planning, thus obtaining the best of both approaches? We show that goal-conditioned policies learned with RL can be incorporated into planning, such that a planner can focus on which states to reach, rather than how those states are reached. However, with complex state observations such as images, not all inputs represent valid states. We therefore also propose using a latent variable model to compactly represent the set of valid states for the planner, such that the policies provide an abstraction of actions, and the latent variable model provides an abstraction of states. We compare our method with planning-based and model-free methods and find that our method significantly outperforms prior work when evaluated on image-based tasks that require non-greedy, multi-staged behavior.
****
Goal-conditioned Imitation Learning
Yiming Ding, Carlos Florensa, Pieter Abbeel, Mariano Phielipp
Designing rewards for Reinforcement Learning (RL) is challenging because it needs to convey the desired task, be efficient to optimize, and be easy to compute. The latter is particularly problematic when applying RL to robotics, where detecting whether the desired configuration is reached might require considerable supervision and instrumentation. Furthermore, we are often interested in being able to reach a wide range of configurations, hence setting up a different reward every time might be unpractical. Methods like Hindsight Experience Replay (HER) have recently shown promise to learn policies able to reach many goals, without the need of a reward. Unfortunately, without tricks like resetting to points along the trajectory, HER might require many samples to discover how to reach certain areas of the state-space. In this work we propose a novel algorithm goalGAIL, which incorporates demonstrations to drastically speed up the convergence to a policy able to reach any goal, surpassing the performance of an agent trained with other Imitation Learning algorithms. Furthermore, we show our method can also be used when the available expert trajectories do not contain the actions or when the expert is suboptimal, which makes it applicable when only kinesthetic, third person or noisy demonstration is available.
****
Superset Technique for Approximate Recovery in One-Bit Compressed Sensing
Larkin Flodin, Venkata Gandikota, Arya Mazumdar
One-bit compressed sensing (1bCS) is a method of signal acquisition under extreme measurement quantization that gives important insights on the limits of signal compression and analog-to-digital conversion. The setting is also equivalent to the problem of learning a sparse hyperplane-classifier. In this paper, we propose a generic approach for signal recovery in nonadaptive 1bCS that leads to improved sample complexity for approximate recovery for a variety of signal models, including nonnegative signals and binary signals. We construct 1bCS matrices that are universal - i.e. work for all signals under a model - and at the same time recover very general random sparse signals with high probability. In our approach, we divide the set of samples (measurements) into two parts, and use the first part to recover the superset of the support of a sparse vector. The second set of measurements is then used to approximate the signal within the superset. While support recovery in 1bCS is well-studied, recovery of superset of the support requires fewer samples, which then leads to an overall reduction in sample complexity for approximate recovery.
****
Iterative Least Trimmed Squares for Mixed Linear Regression
Yanyao Shen, Sujay Sanghavi
Given a linear regression setting, Iterative Least Trimmed Squares (ILTS) involves alternating between (a) selecting the subset of samples with lowest current loss, and (b) re-fitting the linear model only on that subset. Both steps are very fast and simple. In this paper, we analyze ILTS in the setting of mixed linea

r regression with corruptions (MLR-C). We first establish deterministic conditions (on the features etc.) under which the ILTS iterate converges linearly to the closest mixture component. We also provide a global algorithm that uses ILTS as a subroutine, to fully solve mixed linear regressions with corruptions. We then evaluate it for the widely studied setting of isotropic Gaussian features, and establish that we match or better existing results in terms of sample complexity. Finally, we provide an ODE analysis for a gradient-descent variant of ILTS that has optimal time complexity. Our results provide initial theoretical evidence that iteratively fitting to the best subset of samples -- a potentially widely applicable idea -- can provably provide state of the art performance in bad training data settings.
****

Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance

Kimia Nadjahi, Alain Durmus, Umut Simsekli, Roland Badeau

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.
****

Time-series Generative Adversarial Networks

Jinsung Yoon, Daniel Jarrett, Mihaela van der Schaar

A good generative model for time-series data should preserve temporal dynamics, in the sense that new sequences respect the original relationships between variables across time. Existing methods that bring generative adversarial networks (GANs) into the sequential setting do not adequately attend to the temporal correlations unique to time-series data. At the same time, supervised models for sequence prediction - which allow finer control over network dynamics - are inherently deterministic. We propose a novel framework for generating realistic time-series data that combines the flexibility of the unsupervised paradigm with the control afforded by supervised training. Through a learned embedding space jointly optimized with both supervised and adversarial objectives, we encourage the network to adhere to the dynamics of the training data during sampling. Empirically, we evaluate the ability of our method to generate realistic samples using a variety of real and synthetic time-series datasets. Qualitatively and quantitatively, we find that the proposed framework consistently and significantly outperforms state-of-the-art benchmarks with respect to measures of similarity and predictive ability.
****

Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup

Sebastian Goldt, Madhu Advani, Andrew M. Saxe, Florent Krzakala, Lenka Zdeborová

Deep neural networks achieve stellar generalisation even when they have enough parameters to easily fit all their training data. We study this phenomenon by analysing the dynamics and the performance of over-parameterised two-layer neural networks in the teacher-student setup, where one network, the student, is trained on data generated by another network, called the teacher. We show how the dynamics of stochastic gradient descent (SGD) is captured by a set of differential equations and prove that this description is asymptotically exact in the limit of large inputs. Using this framework, we calculate the final generalisation error of student networks that have more parameters than their teachers. We find that the final generalisation error of the student increases with network size when training only the first layer, but stays constant or even decreases with size when training both layers. We show that these different behaviours have their root in the different solutions SGD finds for different activation functions. Our results indicate that achieving good generalisation in neural networks goes beyond the properties of SGD alone and depends on the interplay of at least the algorithm, the model architecture, and the data set.
****

Learning Nonsymmetric Determinantal Point Processes
Mike Gartrell, Victor-Emmanuel Brunel, Elvis Dohmatob, Syrine Krichene
Determinantal point processes (DPPs) have attracted substantial attention as an
elegant probabilistic model that captures the balance between quality and divers
ity within sets. DPPs are conventionally parameterized by a positive semi-defin
ite kernel matrix, and this symmetric kernel encodes only repulsive interactions
 between items. These so-called symmetric DPPs have significant expressive powe
r, and have been successfully applied to a variety of machine learning tasks, in
cluding recommendation systems, information retrieval, and automatic summarizati
on, among many others. Efficient algorithms for learning symmetric DPPs and sam
pling from these models have been reasonably well studied. However, relatively
little attention has been given to nonsymmetric DPPs, which relax the symmetric
constraint on the kernel. Nonsymmetric DPPs allow for both repulsive and attract
ive item interactions, which can significantly improve modeling power, resulting
 in a model that may better fit for some applications. We present a method that
 enables a tractable algorithm, based on maximum likelihood estimation, for lear
ning nonsymmetric DPPs from data composed of observed subsets. Our method impose
s a particular decomposition of the nonsymmetric kernel that enables such tracta
ble learning algorithms, which we analyze both theoretically and experimentally.
  We evaluate our model on synthetic and real-world datasets, demonstrating impr
oved predictive performance compared to symmetric DPPs, which have previously sh
own strong performance on modeling tasks associated with these datasets.
****
Quantum Embedding of Knowledge for Reasoning
Dinesh Garg, Shajith Ikbal, Santosh K. Srivastava, Harit Vishwakarma, Hima Karan
am, L Venkata Subramaniam
Statistical Relational Learning (SRL) methods are the most widely used technique
s to generate distributional representations of the symbolic Knowledge Bases (KB
s). These methods embed any given KB into a vector space by exploiting statistic
al similarities among its entities and predicates but without any guarantee of p
reserving the underlying logical structure of the KB. This, in turn, results in
poor performance of logical reasoning tasks that are solved using such distribut
ional representations. We present a novel approach called Embed2Reason (E2R) tha
t embeds a symbolic KB into a vector space in a logical structure preserving man
ner. This approach is inspired by the theory of Quantum Logic. Such an embedding
 allows answering membership based complex logical reasoning queries with impres
sive accuracy improvements over popular SRL baselines.
****
Online Normalization for Training Neural Networks
Vitaliy Chiley, Ilya Sharapov, Atli Kosson, Urs Koster, Ryan Reece, Sofia Samani
ego de la Fuente, Vishal Subbiah, Michael James
Online Normalization is a new technique for normalizing the hidden activations o
f a neural network. Like Batch Normalization, it normalizes the sample dimension
. While Online Normalization does not use batches, it is as accurate as Batch No
rmalization. We resolve a theoretical limitation of Batch Normalization by intro
ducing an unbiased technique for computing the gradient of normalized activation
s. Online Normalization works with automatic differentiation by adding statistic
al normalization as a primitive. This technique can be used in cases not covered
 by some other normalizers, such as recurrent networks, fully connected networks
, and networks with activation memory requirements prohibitive for batching. We
show its applications to image classification, image segmentation, and language
modeling. We present formal proofs and experimental results on ImageNet, CIFAR,
and PTB datasets.
****
Equitable Stable Matchings in Quadratic Time
Nikolaos Tziavelis, Ioannis Giannakopoulos, Katerina Doka, Nectarios Koziris, Pa
nagiotis Karras
Can a stable matching that achieves high equity among the two sides of a market
be reached in quadratic time? The Deferred Acceptance (DA) algorithm finds a sta
ble matching that is biased in favor of one side; optimizing apt equity measures

 is strongly NP-hard. A proposed approximation algorithm offers a guarantee only with respect to the DA solutions. Recent work introduced Deferred Acceptance with Compensation Chains (DACC), a class of algorithms that can reach any stable matching in O(n^4) time, but did not propose a way to achieve good equity. In this paper, we propose an alternative that is computationally simpler and achieves high equity too. We introduce Monotonic Deferred Acceptance (MDA), a class of algorithms that progresses monotonically towards a stable matching; we couple MDA with a mechanism we call Strongly Deferred Acceptance (SDA), to build an algorithm that reaches an equitable stable matching in quadratic time; we amend this algorithm with a few low-cost local search steps to what we call Deferred Local Search (DLS), and demonstrate experimentally that it outperforms previous solutions in terms of equity measures and matches the most efficient ones in runtime.
****
Making AI Forget You: Data Deletion in Machine Learning
Antonio Ginart, Melody Guan, Gregory Valiant, James Y. Zou
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 New Defense Against Adversarial Images: Turning a Weakness into a Strength
Shengyuan Hu, Tao Yu, Chuan Guo, Wei-Lun Chao, Kilian Q. Weinberger
Natural images are virtually surrounded by low-density misclassified regions that can be efficiently discovered by gradient-guided search --- enabling the generation of adversarial images. While many techniques for detecting these attacks have been proposed, they are easily bypassed when the adversary has full knowledge of the detection mechanism and adapts the attack strategy accordingly. In this paper, we adopt a novel perspective and regard the omnipresence of adversarial perturbations as a strength rather than a weakness. We postulate that if an image has been tampered with, these adversarial directions either become harder to find with gradient methods or have substantially higher density than for natural images. We develop a practical test for this signature characteristic to successfully detect adversarial attacks, achieving unprecedented accuracy under the white-box setting where the adversary is given full knowledge of our detection mechanism.
****
Hamiltonian descent for composite objectives
Brendan O'Donoghue, Chris J. Maddison
In optimization the duality gap between the primal and the dual problems is a measure of the suboptimality of any primal-dual point. In classical mechanics the equations of motion of a system can be derived from the Hamiltonian function, which is a quantity that describes the total energy of the system. In this paper we consider a convex optimization problem consisting of the sum of two convex functions, sometimes referred to as a composite objective, and we identify the duality gap to be the `energy' of the system. In the Hamiltonian formalism the energy is conserved, so we add a contractive term to the standard equations of motion so that this energy decreases linearly (ie, geometrically) with time. This yields a continuous-time ordinary differential equation (ODE) in the primal and dual variables which converges to zero duality gap, ie, optimality. This ODE has several useful properties: it induces a natural operator splitting; at convergence it yields both the primal and dual solutions; and it is invariant to affine transformation despite only using first order information. We provide several discretizations of this ODE, some of which are new algorithms and others correspond to known techniques, such as the alternating direction method of multipliers (ADMM). We conclude with some numerical examples that show the promise of our approach. We give an example where our technique can solve a convex quadratic minimization problem orders of magnitude faster than several commonly-used gradient methods, including conjugate gradient, when the conditioning of the problem is poor. Our framework provides new insights into previously known algorithms in the literature as well as providing a technique to generate new primal-dual algor

ithms.
****
Game Design for Eliciting Distinguishable Behavior
Fan Yang, Liu Leqi, Yifan Wu, Zachary Lipton, Pradeep K. Ravikumar, Tom M. Mitchell, William W. Cohen
The ability to inferring latent psychological traits from human behavior is key to developing personalized human-interacting machine learning systems. Approaches to infer such traits range from surveys to manually-constructed experiments and games. However, these traditional games are limited because they are typically designed based on heuristics. In this paper, we formulate the task of designing behavior diagnostic games that elicit distinguishable behavior as a mutual information maximization problem, which can be solved by optimizing a variational lower bound. Our framework is instantiated by using prospect theory to model varying player traits, and Markov Decision Processes to parameterize the games. We validate our approach empirically, showing that our designed games can successfully distinguish among players with different traits, outperforming manually-designed ones by a large margin.
****
Divergence-Augmented Policy Optimization
Qing Wang, Yingru Li, Jiechao Xiong, Tong Zhang
In deep reinforcement learning, policy optimization methods need to deal with issues such as function approximation and the reuse of off-policy data. Standard policy gradient methods do not handle off-policy data well, leading to premature convergence and instability. This paper introduces a method to stabilize policy optimization when off-policy data are reused. The idea is to include a Bregman divergence between the behavior policy that generates the data and the current policy to ensure small and safe policy updates with off-policy data. The Bregman divergence is calculated between the state distributions of two policies, instead of only on the action probabilities, leading to a divergence augmentation formulation.
Empirical experiments on Atari games show that in the data-scarce scenario where the reuse of off-policy data becomes necessary, our method can achieve better performance than other state-of-the-art deep reinforcement learning algorithms.
****
Gaussian-Based Pooling for Convolutional Neural Networks
Takumi Kobayashi
Convolutional neural networks (CNNs) contain local pooling to effectively downsize feature maps for increasing computation efficiency as well as robustness to input variations. The local pooling methods are generally formulated in a form of convex combination of local neuron activations for retaining the characteristics of an input feature map in a manner similar to image downscaling. In this paper, to improve performance of CNNs, we propose a novel local pooling method based on the Gaussian-based probabilistic model over local neuron activations for flexibly pooling (extracting) features, in contrast to the previous model restricting the output within the convex hull of local neurons. In the proposed method, the local neuron activations are aggregated into the statistics of mean and standard deviation in a Gaussian distribution, and then on the basis of those statistics, we construct the probabilistic model suitable for the pooling in accordance with the knowledge about local pooling in CNNs. Through the probabilistic model equipped with trainable parameters, the proposed method naturally integrates two schemes of adaptively training the pooling form based on input feature maps and stochastically performing the pooling throughout the end-to-end learning. The experimental results on image classification demonstrate that the proposed method favorably improves performance of various CNNs in comparison with the other pooling methods.
****
Band-Limited Gaussian Processes: The Sinc Kernel
Felipe Tobar
We propose a novel class of Gaussian processes (GPs) whose spectra have compact support, meaning that their sample trajectories are almost-surely band limited.

As a complement to the growing literature on spectral design of covariance kernels, the core of our proposal is to model power spectral densities through a rectangular function, which results in a kernel based on the sinc function with straightforward extensions to non-centred (around zero frequency) and frequency-varying cases. In addition to its use in regression, the relationship between the sinc kernel and the classic theory is illuminated, in particular, the Shannon-Nyquist theorem is interpreted as posterior reconstruction under the proposed kernel. Additionally, we show that the sinc kernel is instrumental in two fundamental signal processing applications: first, in stereo amplitude modulation, where the non-centred sinc kernel arises naturally. Second, for band-pass filtering, where the proposed kernel allows for a Bayesian treatment that is robust to observation noise and missing data. The developed theory is complemented with illustrative graphic examples and validated experimentally using real-world data.
****
Break the Ceiling: Stronger Multi-scale Deep Graph Convolutional Networks
Sitao Luan, Mingde Zhao, Xiao-Wen Chang, Doina Precup
Recently, neural network based approaches have achieved significant progress for solving large, complex, graph-structured problems. Nevertheless, the advantages of multi-scale information and deep architectures have not been sufficiently exploited. In this paper, we first analyze key factors constraining the expressive power of existing Graph Convolutional Networks (GCNs), including the activation function and shallow learning mechanisms. Then, we generalize spectral graph convolution and deep GCN in block Krylov subspace forms, upon which we devise two architectures, both scalable in depth however making use of multi-scale information differently. On several node classification tasks, the proposed architectures achieve state-of-the-art performance.
****
Bayesian Optimization with Unknown Search Space
Huong Ha, Santu Rana, Sunil Gupta, Thanh Nguyen, Hung Tran-The, Svetha Venkatesh
Applying Bayesian optimization in problems wherein the search space is unknown is challenging. To address this problem, we propose a systematic volume expansion strategy for the Bayesian optimization. We devise a strategy to guarantee that in iterative expansions of the search space, our method can find a point whose function value within epsilon of the objective function maximum. Without the need to specify any parameters, our algorithm automatically triggers a minimal expansion required iteratively. We derive analytic expressions for when to trigger the expansion and by how much to expand. We also provide theoretical analysis to show that our method achieves epsilon-accuracy after a finite number of iterations. We demonstrate our method on both benchmark test functions and machine learning hyper-parameter tuning tasks and demonstrate that our method outperforms baselines.
****
Towards closing the gap between the theory and practice of SVRG
Othmane Sebbouh, Nidham Gazagnadou, Samy Jelassi, Francis Bach, Robert Gower
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 Unifying Framework for Spectrum-Preserving Graph Sparsification and Coarsening
Gecia Bravo Hermsdorff, Lee Gunderson
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.
****
Error Correcting Output Codes Improve Probability Estimation and Adversarial Robustness of Deep Neural Networks
Gunjan Verma, Ananthram Swami
Modern machine learning systems are susceptible to adversarial examples; inputs

which clearly preserve the characteristic semantics of a given class, but whose classification is (usually confidently) incorrect. Existing approaches to adversarial
defense generally rely on modifying the input, e.g. quantization, or the learned model parameters, e.g. via adversarial training. However, recent research has shown that most such approaches succumb to adversarial examples when different norms or more sophisticated adaptive attacks are considered. In this paper, we propose a fundamentally different approach which instead changes the way the output is represented and decoded. This simple approach achieves state-of-the-art robustness to adversarial examples for L 2 and L ∞ based adversarial perturbations on MNIST and CIFAR10. In addition, even under strong white-box attacks, we find that our model often assigns adversarial examples a low probability; those with high probability are usually interpretable, i.e. perturbed towards the perceptual boundary between the original and adversarial class. Our approach has several advantages: it yields more meaningful probability estimates, is extremely fast during training and testing, requires essentially no architectural changes to existing discriminative learning pipelines, is wholly complementary to other defense approaches including adversarial training, and does not sacrifice benign test set performance
****

KerGM: Kernelized Graph Matching

Zhen Zhang, Yijian Xiang, Lingfei Wu, Bing Xue, Arye Nehorai

Graph matching plays a central role in such fields as computer vision, pattern recognition, and bioinformatics. Graph matching problems can be cast as two types of quadratic assignment problems (QAPs): Koopmans-Beckmann's QAP or Lawler's QAP. In our paper, we provide a unifying view for these two problems by introducing new rules for array operations in Hilbert spaces. Consequently, Lawler's QAP can be considered as the Koopmans-Beckmann's alignment between two arrays in reproducing kernel Hilbert spaces (RKHS), making it possible to efficiently solve the problem without computing a huge affinity matrix. Furthermore, we develop the entropy-regularized Frank-Wolfe (EnFW) algorithm for optimizing QAPs, which has the same convergence rate as the original FW algorithm while dramatically reducing the computational burden for each outer iteration. We conduct extensive experiments to evaluate our approach, and show that our algorithm significantly outperforms the state-of-the-art in both matching accuracy and scalability.
****

On Human-Aligned Risk Minimization

Liu Leqi, Adarsh Prasad, Pradeep K. Ravikumar

The statistical decision theoretic foundations of modern machine learning have largely focused on the minimization of the expectation of some loss function for a given task. However, seminal results in behavioral economics have shown that human decision-making is based on different risk measures than the expectation of any given loss function. In this paper, we pose the following simple question: in contrast to minimizing expected loss, could we minimize a better human-aligned risk measure? While this might not seem natural at first glance, we analyze the properties of such a revised risk measure, and surprisingly show that it might also better align with additional desiderata like fairness that have attracted considerable recent attention. We focus in particular on a class of human-aligned risk measures inspired by cumulative prospect theory. We empirically study these risk measures, and demonstrate their improved performance on desiderata such as fairness, in contrast to the traditional workhorse of expected loss minimization.
****

Robustness Verification of Tree-based Models

Hongge Chen, Huan Zhang, Si Si, Yang Li, Duane Boning, Cho-Jui Hsieh

We study the robustness verification problem of tree based models, including random forest (RF) and gradient boosted decision tree (GBDT).
Formal robustness verification of decision tree ensembles involves finding the exact minimal adversarial perturbation or a guaranteed lower bound of it. Existing approaches cast this verification problem into a mixed integer linear programm

ing (MILP) problem, which finds the minimal adversarial distortion in exponentia
l time so is impractical for large ensembles. Although this verification problem
is NP-complete in general, we give a more precise complexity characterization.
We show that there is a simple linear time algorithm for verifying a single tree
, and for tree ensembles the verification problem can be cast as a max-clique pr
oblem on a multi-partite boxicity graph. For low dimensional problems when boxic
ity can be viewed as constant, this reformulation leads to a polynomial time alg
orithm. For general problems, by exploiting the boxicity of the graph, we devise
an efficient verification algorithm that can give tight lower bounds on robustn
ess of decision tree ensembles, and allows iterative improvement and any-time te
rmination. On RF/GBDT models trained on a variety of datasets, we significantly
outperform the lower bounds obtained by relaxing the MILP formulation into a lin
ear program (LP), and are hundreds times faster than solving MILPs to get the ex
act minimal adversarial distortion. Our proposed method is capable of giving tig
ht robustness verification bounds on large GBDTs with hundreds of deep trees.
****

Provable Non-linear Inductive Matrix Completion
Kai Zhong, Zhao Song, Prateek Jain, Inderjit S. Dhillon
Consider a standard recommendation/retrieval problem where given a query, the go
al is to retrieve the most relevant items. Inductive matrix completion (IMC) met
hod is a standard approach for this problem where the given query as well as the
items are embedded in a common low-dimensional space. The inner product between
a query embedding and an item embedding reflects relevance of the (query, item)
pair. Non-linear IMC (NIMC) uses non-linear networks to embed the query as wel
l as items, and is known to be highly effective for a variety of tasks, such as
video recommendations for users, semantic web search, etc. Despite its wide usag
e, existing literature lacks rigorous understanding of NIMC models. A key challe
nge in analyzing such models is to deal with the non-convexity arising out of no
n-linear embeddings in addition to the non-convexity arising out of the low-dime
nsional restriction of the embedding space, which is akin to the low-rank restri
ction in the standard matrix completion problem. In this paper, we provide the
first theoretical analysis for a simple NIMC model in the realizable setting, w
here the relevance score of a (query, item) pair is formulated as the inner prod
uct between their single-layer neural representations. Our results show that und
er mild assumptions we can recover the ground truth parameters of the NIMC model
using standard (stochastic) gradient descent methods if the methods are initial
ized within a small distance to the optimal parameters. We show that a standard
tensor method can be used to initialize the solution within the required distanc
e to the optimal parameters. Furthermore, we show that the number of query-item
relevance observations required, a key parameter in learning such models, scales
nearly linearly with the input dimensionality thus matching existing results fo
r the standard linear inductive matrix completion.
****

STAR-Caps: Capsule Networks with Straight-Through Attentive Routing
Karim Ahmed, Lorenzo Torresani
Capsule networks have been shown to be powerful models for image classification,
thanks to their ability to represent and capture viewpoint variations of an obj
ect. However, the high computational complexity of capsule networks that stems f
rom the recurrent dynamic routing poses a major drawback making their use for la
rge-scale image classification challenging. In this work, we propose Star-Caps a
capsule-based network that exploits a straight-through attentive routing to add
ress the drawbacks of capsule networks. By utilizing attention modules augmented
by differentiable binary routers, the proposed mechanism estimates the routing
coefficients between capsules without recurrence, as opposed to prior related wo
rk. Subsequently, the routers utilize straight-through estimators to make binary
decisions to either connect or disconnect the route between capsules, allowing
stable and faster performance. The experiments conducted on several image classi
fication datasets, including MNIST, SmallNorb, CIFAR-10, CIFAR-100, and ImageNet
show that Star-Caps outperforms the baseline capsule networks.
****

Self-attention with Functional Time Representation Learning
Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, Kannan Achan
Sequential modelling with self-attention has achieved cutting edge performances
in natural language processing. With advantages in model flexibility, computatio
n complexity and interpretability, self-attention is gradually becoming a key co
mponent in event sequence models. However, like most other sequence models, self
-attention does not account for the time span between events and thus captures s
equential signals rather than temporal patterns.
Without relying on recurrent network structures, self-attention recognizes event
 orderings via positional encoding. To bridge the gap between modelling time-ind
ependent and time-dependent event sequence, we introduce a functional feature ma
p that embeds time span into high-dimensional spaces. By constructing the associ
ated translation-invariant time kernel function, we reveal the functional forms
of the feature map under classic functional function analysis results, namely Bo
chner's Theorem and Mercer's Theorem. We propose several models to learn the fun
ctional time representation and the interactions with event representation. Thes
e methods are evaluated on real-world datasets under various continuous-time eve
nt sequence prediction tasks. The experiments reveal that the proposed methods c
ompare favorably to baseline models while also capture useful time-event interac
tions.
****
Multi-label Co-regularization for Semi-supervised Facial Action Unit Recognition
Xuesong Niu, Hu Han, Shiguang Shan, Xilin Chen
Facial action units (AUs) recognition is essential for emotion analysis and has
been widely applied in mental state analysis. Existing work on AU recognition us
ually requires big face dataset with accurate AU labels. However, manual AU anno
tation requires expertise and can be time-consuming. In this work, we propose a
semi-supervised approach for AU recognition utilizing a large number of web face
 images without AU labels and a small face dataset with AU labels inspired by th
e co-training methods. Unlike traditional co-training methods that require provi
ded multi-view features and model re-training, we propose a novel co-training me
thod, namely multi-label co-regularization, for semi-supervised facial AU recogn
ition. Two deep neural networks are used to generate multi-view features for bot
h labeled and unlabeled face images, and a multi-view loss is designed to enforc
e the generated features from the two views to be conditionally independent repr
esentations. In order to obtain consistent predictions from the two views, we fu
rther design a multi-label co-regularization loss aiming to minimize the distanc
e between the predicted AU probability distributions of the two views. In additi
on, prior knowledge of the relationship between individual AUs is embedded throu
gh a graph convolutional network (GCN) for exploiting useful information from th
e big unlabeled dataset. Experiments on several benchmarks show that the propose
d approach can effectively leverage large datasets of unlabeled face images to i
mprove the AU recognition robustness and outperform the state-of-the-art semi-su
pervised AU recognition methods.
****
A Primal Dual Formulation For Deep Learning With Constraints
Yatin Nandwani, Abhishek Pathak, Mausam, Parag Singla
For several problems of interest, there are natural constraints which exist over
 the output label space. For example, for the joint task of NER and POS labeling
, these constraints might specify that the NER label ‘organization’ is consisten
t only with the POS labels ‘noun’ and ‘preposition’. These constraints can be a
great way of injecting prior knowledge into a deep learning model, thereby impro
ving overall performance. In this paper, we present a constrained optimization f
ormulation for training a deep network with a given set of hard constraints on o
utput labels. Our novel approach first converts the label constraints into soft
logic constraints over probability distributions outputted by the network. It th
en converts the constrained optimization problem into an alternating min-max opt
imization with Lagrangian variables defined for each constraint. Since the cons
traints are independent of the target labels, our framework easily generalizes
to semi-supervised setting. We experiment on the tasks of Semantic Role Labeling

(SRL), Named Entity Recognition (NER) tagging, and fine-grained entity typing and show that our constraints not only significantly reduce the number of constraint violations, but can also result in state-of-the-art performance
****
DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections
Ofir Nachum, Yinlam Chow, Bo Dai, Lihong Li
In many real-world reinforcement learning applications, access to the environment is limited to a fixed dataset, instead of direct (online) interaction with the environment. When using this data for either evaluation or training of a new policy, accurate estimates of discounted stationary distribution ratios -- correction terms which quantify the likelihood that the new policy will experience a certain state-action pair normalized by the probability with which the state-action pair appears in the dataset -- can improve accuracy and performance. In this work, we propose an algorithm, DualDICE, for estimating these quantities. In contrast to previous approaches, our algorithm is agnostic to knowledge of the behavior policy (or policies) used to generate the dataset. Furthermore, our algorithm eschews any direct use of importance weights, thus avoiding potential optimization instabilities endemic of previous methods. In addition to providing theoretical guarantees, we present an empirical study of our algorithm applied to off-policy policy evaluation and find that our algorithm significantly improves accuracy compared to existing techniques.
****
Generalization Bounds of Stochastic Gradient Descent for Wide and Deep Neural Networks
Yuan Cao, 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.
****
Intrinsic dimension of data representations in deep neural networks
Alessio Ansuini, Alessandro Laio, Jakob H. Macke, Davide Zoccolan
Deep neural networks progressively transform their inputs across multiple processing layers. What are the geometrical properties of the representations learned by these networks? Here we study the intrinsic dimensionality (ID) of data representations, i.e. the minimal number of parameters needed to describe a representation. We find that, in a trained network, the ID is orders of magnitude smaller than the number of units in each layer. Across layers, the ID first increases and then progressively decreases in the final layers. Remarkably, the ID of the last hidden layer predicts classification accuracy on the test set. These results can neither be found by linear dimensionality estimates (e.g., with principal component analysis), nor in representations that had been artificially linearized. They are neither found in untrained networks, nor in networks that are trained on randomized labels. This suggests that neural networks that can generalize are those that transform the data into low-dimensional, but not necessarily flat manifolds.
****
Program Synthesis and Semantic Parsing with Learned Code Idioms
Eui Chul Shin, Miltiadis Allamanis, Marc Brockschmidt, Alex Polozov
Program synthesis of general-purpose source code from natural language specifications is challenging due to the need to reason about high-level patterns in the target program and low-level implementation details at the same time. In this work, we present Patois, a system that allows a neural program synthesizer to explicitly interleave high-level and low-level reasoning at every generation step. It accomplishes this by automatically mining common code idioms from a given corpus, incorporating them into the underlying language for neural synthesis, and training a tree-based neural synthesizer to use these idioms during code generation. We evaluate Patois on two complex semantic parsing datasets and show that using learned code idioms improves the synthesizer's accuracy.

****
Data-driven Estimation of Sinusoid Frequencies
Gautier Izacard, Sreyas Mohan, Carlos Fernandez-Granda
Frequency estimation is a fundamental problem in signal processing, with applica tions in radar imaging, underwater acoustics, seismic imaging, and spectroscopy. The goal is to estimate the frequency of each component in a multisinusoidal si gnal from a finite number of noisy samples. A recent machine-learning approach u ses a neural network to output a learned representation with local maxima at the position of the frequency estimates. In this work, we propose a novel neural-ne twork architecture that produces a significantly more accurate representation, a nd combine it with an additional neural-network module trained to detect the num ber of frequencies. This yields a fast, fully-automatic method for frequency est imation that achieves state-of-the-art results. In particular, it outperforms ex isting techniques by a substantial margin at medium-to-high noise levels.
****
Discovering Neural Wirings
Mitchell Wortsman, Ali Farhadi, Mohammad Rastegari
The success of neural networks has driven a shift in focus from feature engineer ing to architecture engineering. However, successful networks today are construc ted using a small and manually defined set of building blocks. Even in methods o f neural architecture search (NAS) the network connectivity patterns are largely constrained. In this work we propose a method for discovering neural wirings. W e relax the typical notion of layers and instead enable channels to form connect ions independent of each other. This allows for a much larger space of possible networks. The wiring of our network is not fixed during training -- as we learn the network parameters we also learn the structure itself. Our experiments demon strate that our learned connectivity outperforms hand engineered and randomly wi red networks. By learning the connectivity of MobileNetV1we boost the ImageNet a ccuracy by 10% at ~41M FLOPs. Moreover, we show that our method generalizes to r ecurrent and continuous time networks.
Our work may also be regarded as unifying core aspects of the neural architectur e search problem with sparse neural network learning. As NAS becomes more fine g rained, finding a good architecture is akin to finding a sparse subnetwork of th e complete graph. Accordingly, DNW provides an effective mechanism for discoveri ng sparse subnetworks of predefined architectures in a single training run. Thou gh we only ever use a small percentage of the weights during the forward pass, w e still play the so-called initialization lottery with a combinatorial number of subnetworks. Code and pretrained models are available at https://github.com/all enai/dnw while additional visualizations may be found at https://mitchellnw.gith ub.io/blog/2019/dnw/.
****
Locally Private Learning without Interaction Requires Separation
Amit Daniely, Vitaly Feldman
We consider learning under the constraint of local differential privacy (LDP). F or many learning problems known efficient algorithms in this model require many rounds of communication between the server and the clients holding the data poin ts. Yet multi-round protocols are prohibitively slow in practice due to network latency and, as a result, currently deployed large-scale systems are limited to a single round. Despite significant research interest, very little is known abou t which learning problems can be solved by such non-interactive systems. The onl y lower bound we are aware of is for PAC learning an artificial class of functio ns with respect to a uniform distribution (Kasiviswanathan et al., 2008).
****
Fixing the train-test resolution discrepancy
Hugo Touvron, Andrea Vedaldi, Matthijs Douze, Herve Jegou
Data-augmentation is key to the training of neural networks for image classifica tion. This paper first shows that existing augmentations induce a significant di screpancy between the size of the objects seen by the classifier at train and te st time: in fact, a lower train resolution improves the classification at test t ime!

****
Quadratic Video Interpolation
Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, Ming-Hsuan Yang
Video interpolation is an important problem in computer vision, which helps over come the temporal limitation of camera sensors. Existing video interpolation met hods usually assume uniform motion between consecutive frames and use linear mod els for interpolation, which cannot well approximate the complex motion in the r eal world. To address these issues, we propose a quadratic video interpolation m ethod which exploits the acceleration information in videos. This method allows prediction with curvilinear trajectory and variable velocity, and generates more accurate interpolation results. For high-quality frame synthesis, we develop a flow reversal layer to estimate flow fields starting from the unknown target fra me to the source frame. In addition, we present techniques for flow refinement. Extensive experiments demonstrate that our approach performs favorably against t he existing linear models on a wide variety of video datasets.
****
Self-supervised GAN: Analysis and Improvement with Multi-class Minimax Game
Ngoc-Trung Tran, Viet-Hung Tran, Bao-Ngoc Nguyen, Linxiao Yang, Ngai-Man (Man) C heung
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.
****
Learning step sizes for unfolded sparse coding
Pierre Ablin, Thomas Moreau, Mathurin Massias, Alexandre Gramfort
Sparse coding is typically solved by iterative optimization techniques, such as the Iterative Shrinkage-Thresholding Algorithm (ISTA). Unfolding and learning we ights of ISTA using neural networks is a practical way to accelerate estimation. In this paper, we study the selection of adapted step sizes for ISTA. We show t hat a simple step size strategy can improve the convergence rate of ISTA by leve raging the sparsity of the iterates. However, it is impractical in most large-sc ale applications. Therefore, we propose a network architecture where only the st ep sizes of ISTA are learned. We demonstrate that for a large class of unfolded algorithms, if the algorithm converges to the solution of the Lasso, its last la yers correspond to ISTA with learned step sizes. Experiments show that our metho d is competitive with state-of-the-art networks when the solutions are sparse en ough.
****
Efficient Graph Generation with Graph Recurrent Attention Networks
Renjie Liao, Yujia Li, Yang Song, Shenlong Wang, Will Hamilton, David K. Duvenau d, Raquel Urtasun, Richard Zemel
We propose a new family of efficient and expressive deep generative models of gr aphs, called Graph Recurrent Attention Networks (GRANs).
Our model generates graphs one block of nodes and associated edges at a time.
The block size and sampling stride allow us to trade off sample quality for effi ciency.
Compared to previous RNN-based graph generative models, our framework better cap tures the auto-regressive conditioning between the already-generated and to-be-g enerated parts of the graph using Graph Neural Networks (GNNs) with attention.
This not only reduces the dependency on node ordering but also bypasses the long -term bottleneck caused by the sequential nature of RNNs.
Moreover, we parameterize the output distribution per block using a mixture of B ernoulli, which captures the correlations among generated edges within the block .
Finally, we propose to handle node orderings in generation by marginalizing over a family of canonical orderings.
On standard benchmarks, we achieve state-of-the-art time efficiency and sample q uality compared to previous models.
Additionally, we show our model is capable of generating large graphs of up to 5

K nodes with good quality.
Our code is released at: \url{https://github.com/lrjconan/GRAN}.
****
Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks
Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, Silvio Savarese
Predicting the future trajectories of multiple interacting pedestrians in a scene has become an increasingly important problem for many different applications ranging from control of autonomous vehicles and social robots to security and surveillance. This problem is compounded by the presence of social interactions between humans and their physical interactions with the scene. While the existing literature has explored some of these cues, they mainly ignored the multimodal nature of each human's future trajectory which is noticeably influenced by the intricate social interactions. In this paper, we present Social-BiGAT, a graph-based generative adversarial network that generates realistic, multimodal trajectory predictions for multiple pedestrians in a scene. Our method is based on a graph attention network (GAT) that learns feature representations that encode the social interactions between humans in the scene, and a recurrent encoder-decoder architecture that is trained adversarially to predict, based on the features, the humans' paths. We explicitly account for the multimodal nature of the prediction problem by forming a reversible transformation between each scene and its latent noise vector, as in Bicycle-GAN. We show that our framework achieves state-of-the-art performance comparing it to several baselines on existing trajectory forecasting benchmarks.
****
Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds
Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, Niki Trigoni
We propose a novel, conceptually simple and general framework for instance segmentation on 3D point clouds. Our method, called 3D-BoNet, follows the simple design philosophy of per-point multilayer perceptrons (MLPs). The framework directly regresses 3D bounding boxes for all instances in a point cloud, while simultaneously predicting a point-level mask for each instance. It consists of a backbone network followed by two parallel network branches for 1) bounding box regression and 2) point mask prediction. 3D-BoNet is single-stage, anchor-free and end-to-end trainable. Moreover, it is remarkably computationally efficient as, unlike existing approaches, it does not require any post-processing steps such as non-maximum suppression, feature sampling, clustering or voting. Extensive experiments show that our approach surpasses existing work on both ScanNet and S3DIS datasets while being approximately 10x more computationally efficient. Comprehensive ablation studies demonstrate the effectiveness of our design.
****
Re-examination of the Role of Latent Variables in Sequence Modeling
Guokun Lai, Zihang Dai, Yiming Yang, Shinjae Yoo
With latent variables, stochastic recurrent models have achieved state-of-the-art performance in modeling sound-wave sequence.
However, opposite results are also observed in other domains, where standard recurrent networks often outperform stochastic models.
To better understand this discrepancy, we re-examine the roles of latent variables in stochastic recurrent models for speech density estimation.
Our analysis reveals that under the restriction of fully factorized output distribution in previous evaluations, the stochastic variants were implicitly leveraging intra-step correlation but the deterministic recurrent baselines were prohibited to do so, resulting in an unfair comparison.
To correct the unfairness, we remove such restriction in our re-examination, where all the models can explicitly leverage intra-step correlation with an auto-regressive structure.
Over a diverse set of univariate and multivariate sequential data, including human speech, MIDI music, handwriting trajectory, and frame-permuted speech, our re

sults show that stochastic recurrent models fail to deliver the performance advantage claimed in previous work.
%exhibit any practical advantage despite the claimed theoretical superiority.
In contrast, standard recurrent models equipped with an auto-regressive output distribution consistently perform better, dramatically advancing the state-of-the-art results on three speech datasets.
****
Consistency-based Semi-supervised Learning for Object detection
Jisoo Jeong, Seungeui Lee, Jeesoo Kim, Nojun Kwak
Making a precise annotation in a large dataset is crucial to the performance of object detection. While the object detection task requires a huge number of annotated samples to guarantee its performance, placing bounding boxes for every object in each sample is time-consuming and costs a lot. To alleviate this problem, we propose a Consistency-based Semi-supervised learning method for object Detection (CSD), which is a way of using consistency constraints as a tool for enhancing detection performance by making full use of available unlabeled data. Specifically, the consistency constraint is applied not only for object classification but also for the localization. We also proposed Background Elimination (BE) to avoid the negative effect of the predominant backgrounds on the detection performance. We have evaluated the proposed CSD both in single-stage and two-stage detectors and the results show the effectiveness of our method.
****
Kernel Truncated Randomized Ridge Regression: Optimal Rates and Low Noise Acceleration
Kwang-Sung Jun, Ashok Cutkosky, Francesco Orabona
In this paper we consider the nonparametric least square regression in a Reproducing Kernel Hilbert Space (RKHS). We propose a new randomized algorithm that has optimal generalization error bounds with respect to the square loss, closing a long-standing gap between upper and lower bounds. Moreover, we show that our algorithm has faster finite-time and asymptotic rates on problems where the Bayes risk with respect to the square loss is small. We state our results using standard tools from the theory of least square regression in RKHSs, namely, the decay of the eigenvalues of the associated integral operator and the complexity of the optimal predictor measured through the integral operator.
****
Bandits with Feedback Graphs and Switching Costs
Raman Arora, Teodor Vanislavov Marinov, Mehryar Mohri
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 Combinatorial Optimization with Graph Convolutional Neural Networks
Maxime Gasse, Didier Chetelat, Nicola Ferroni, Laurent Charlin, Andrea Lodi
Combinatorial optimization problems are typically tackled by the branch-and-bound paradigm. We propose a new graph convolutional neural network model for learning branch-and-bound variable selection policies, which leverages the natural variable-constraint bipartite graph representation of mixed-integer linear programs. We train our model via imitation learning from the strong branching expert rule, and demonstrate on a series of hard problems that our approach produces policies that improve upon state-of-the-art machine-learning methods for branching and generalize to instances significantly larger than seen during training. Moreover, we improve for the first time over expert-designed branching rules implemented in a state-of-the-art solver on large problems. Code for reproducing all the experiments can be found at https://github.com/ds4dm/learn2branch.
****
Comparing Unsupervised Word Translation Methods Step by Step
Mareike Hartmann, Yova Kementchedjhieva, Anders Søgaard
Cross-lingual word vector space alignment is the task of mapping the vocabularies of two languages into a shared semantic space, which can be used for dictionar

y induction, unsupervised machine translation, and transfer learning. In the uns
upervised regime, an initial seed dictionary is learned in the absence of any kn
own correspondences between words, through {\bf distribution matching}, and the
seed dictionary is then used to supervise the induction of the final alignment i
n what is typically referred to as a (possibly iterative) {\bf refinement} step.
 We focus on the first step and compare distribution matching techniques in the
 context of language pairs for which mixed training stability and evaluation sco
res have been reported. We show that, surprisingly, when looking at this initia
l step in isolation, vanilla GANs are superior to more recent methods, both in t
erms of precision and robustness. The improvements reported by more recent metho
ds thus stem from the refinement techniques, and we show that we can obtain stat
e-of-the-art performance combining vanilla GANs with such refinement techniques.
****
Learn, Imagine and Create: Text-to-Image Generation from Prior Knowledge
Tingting Qiao, Jing Zhang, Duanqing Xu, Dacheng Tao
Text-to-image generation, i.e. generating an image given a text description, is
a very challenging task due to the significant semantic gap between the two doma
ins. Humans, however, tackle this problem intelligently. We learn from diverse o
bjects to form a solid prior about semantics, textures, colors, shapes, and layo
uts. Given a text description, we immediately imagine an overall visual impressi
on using this prior and, based on this, we draw a picture by progressively addin
g more and more details. In this paper, and inspired by this process, we propose
 a novel text-to-image method called LeicaGAN to combine the above three phases
in a unified framework. First, we formulate the multiple priors learning phase a
s a textual-visual co-embedding (TVE) comprising a text-image encoder for learni
ng semantic, texture, and color priors and a text-mask encoder for learning shap
e and layout priors. Then, we formulate the imagination phase as multiple priors
 aggregation (MPA) by combining these complementary priors and adding noise for
diversity. Lastly, we formulate the creation phase by using a cascaded attentive
 generator (CAG) to progressively draw a picture from coarse to fine. We leverag
e adversarial learning for LeicaGAN to enforce semantic consistency and visual r
ealism. Thorough experiments on two public benchmark datasets demonstrate LeicaG
AN's superiority over the baseline method. Code has been made available at https
://github.com/qiaott/LeicaGAN.
****
Compiler Auto-Vectorization with Imitation Learning
Charith Mendis, Cambridge Yang, Yewen Pu, Dr.Saman Amarasinghe, Michael Carbin
Modern microprocessors are equipped with single instruction multiple data (SIMD)
 or vector instruction sets which allow compilers to exploit fine-grained data l
evel parallelism. To exploit this parallelism, compilers employ auto-vectorizati
on techniques to automatically convert scalar code into vector code. Larsen & Am
arasinghe (2000) first introduced superword level parallelism (SLP) based vector
ization, which is one form of vectorization popularly used by compilers. Current
 compilers employ hand-crafted heuristics and typically only follow one SLP vect
orization strategy which can be suboptimal. Recently, Mendis & Amarasinghe (2018
) formulated the instruction packing problem of SLP vectorization by leveraging
an integer linear programming (ILP) solver, achieving superior runtime performan
ce. In this work, we explore whether it is feasible to imitate optimal decisions
 made by their ILP solution by fitting a graph neural network policy. We show th
at the learnt policy produces a vectorization scheme which is better than indust
ry standard compiler heuristics both in terms of static measures and runtime per
formance. More specifically, the learnt agent produces a vectorization scheme wh
ich has a 22.6% higher average reduction in cost compared to LLVM compiler when
measured using its own cost model and achieves a geometric mean runtime speedup
of 1.015× on the NAS benchmark suite when compared to LLVM’s SLP vectorizer.
****
Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification and Local C
omputations
Debraj Basu, Deepesh Data, Can Karakus, Suhas Diggavi
Communication bottleneck has been identified as a significant issue in distribut

ed optimization of large-scale learning models. Recently, several approaches to mitigate this problem have been proposed, including different forms of gradient compression or computing local models and mixing them iteratively. In this paper we propose Qsparse-local-SGD algorithm, which combines aggressive sparsification with quantization and local computation along with error compensation, by keeping track of the difference between the true and compressed gradients. We propose both synchronous and asynchronous implementations of Qsparse-local-SGD. We analyze convergence for Qsparse-local-SGD in the distributed case, for smooth non-convex and convex objective functions. We demonstrate that Qsparse-local-SGD converges at the same rate as vanilla distributed SGD for many important classes of sparsifiers and quantizers. We use Qsparse-local-SGD to train ResNet-50 on ImageNet, and show that it results in significant savings over the state-of-the-art, in the number of bits transmitted to reach target accuracy.
****

Fast Sparse Group Lasso
Yasutoshi Ida, Yasuhiro Fujiwara, Hisashi Kashima
Sparse Group Lasso is a method of linear regression analysis that finds sparse parameters in terms of both feature groups and individual features.
Block Coordinate Descent is a standard approach to obtain the parameters of Sparse Group Lasso, and iteratively updates the parameters for each parameter group.
However, as an update of only one parameter group depends on all the parameter groups or data points, the computation cost is high when the number of the parameters or data points is large.
This paper proposes a fast Block Coordinate Descent for Sparse Group Lasso.
It efficiently skips the updates of the groups whose parameters must be zeros by using the parameters in one group.
In addition, it preferentially updates parameters in a candidate group set, which contains groups whose parameters must not be zeros.
Theoretically, our approach guarantees the same results as the original Block Coordinate Descent.
Experiments show that our algorithm enhances the efficiency of the original algorithm without any loss of accuracy.
****

Deep Random Splines for Point Process Intensity Estimation of Neural Population Data
Gabriel Loaiza-Ganem, Sean Perkins, Karen Schroeder, Mark Churchland, John P. Cunningham
Gaussian processes are the leading class of distributions on random functions, but they suffer from well known issues including difficulty scaling and inflexibility with respect to certain shape constraints (such as nonnegativity). Here we propose Deep Random Splines, a flexible class of random functions obtained by transforming Gaussian noise through a deep neural network whose output are the parameters of a spline. Unlike Gaussian processes, Deep Random Splines allow us to readily enforce shape constraints while inheriting the richness and tractability of deep generative models. We also present an observational model for point process data which uses Deep Random Splines to model the intensity function of each point process and apply it to neural population data to obtain a low-dimensional representation of spiking activity. Inference is performed via a variational autoencoder that uses a novel recurrent encoder architecture that can handle multiple point processes as input. We use a newly collected dataset where a primate completes a pedaling task, and observe better dimensionality reduction with our model than with competing alternatives.
****

Fast Decomposable Submodular Function Minimization using Constrained Total Variation
Senanayak Sesh Kumar Karri, Francis Bach, Thomas Pock
We consider the problem of minimizing the sum of submodular set functions assuming minimization oracles of each summand function. Most existing approaches reformulate the problem as the convex minimization of the sum of the corresponding Lov\'asz extensions and the squared Euclidean norm, leading to algorithms requirin

g total variation oracles of the summand functions; without further assumptions, these more complex oracles require many calls to the simpler minimization oracles often available in practice. In this paper, we consider a modified convex problem requiring constrained version of the total variation oracles that can be solved with significantly fewer calls to the simple minimization oracles. We support our claims by showing results on graph cuts for 2D and 3D graphs.
****
Deep Signature Transforms
Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, Terry Lyons
The signature is an infinite graded sequence of statistics known to characterise a stream of data up to a negligible equivalence class. It is a transform which has previously been treated as a fixed feature transformation, on top of which a model may be built. We propose a novel approach which combines the advantages of the signature transform with modern deep learning frameworks. By learning an augmentation of the stream prior to the signature transform, the terms of the signature may be selected in a data-dependent way. More generally, we describe how the signature transform may be used as a layer anywhere within a neural network. In this context it may be interpreted as a pooling operation. We present the results of empirical experiments to back up the theoretical justification. Code available at \texttt{github.com/patrick-kidger/Deep-Signature-Transforms}.
****
ResNets Ensemble via the Feynman-Kac Formalism to Improve Natural and Robust Accuracies
Bao Wang, Zuoqiang Shi, Stanley Osher
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.
****
Guided Meta-Policy Search
Russell Mendonca, Abhishek Gupta, Rosen Kralev, Pieter Abbeel, Sergey Levine, Chelsea Finn
Reinforcement learning (RL) algorithms have demonstrated promising results on complex tasks, yet often require impractical numbers of samples because they learn from scratch. Meta-RL aims to address this challenge by leveraging experience from previous tasks so as to more quickly solve new tasks. However, in practice, these algorithms generally also require large amounts of on-policy experience during the \emph{meta-training} process, making them impractical for use in many problems. To this end, we propose to learn a reinforcement learning procedure in a federated way, where individual off-policy learners can solve the individual meta-training tasks, and then consolidate these solutions into a single meta-learner. Since the central meta-learner learns by imitating the solutions to the individual tasks, it can accommodate either the standard meta-RL problem setting, or a hybrid setting where some or all tasks are provided with example demonstrations. The former results in an approach that can leverage policies learned for previous tasks without significant amounts of on-policy data during meta-training, whereas the latter is particularly useful in cases where demonstrations are easy for a person to provide. Across a number of continuous control meta-RL problems, we demonstrate significant improvements in meta-RL sample efficiency in comparison to prior work as well as the ability to scale to domains with visual observations.
****
Learning elementary structures for 3D shape generation and matching
Theo Deprelle, Thibault Groueix, Matthew Fisher, Vladimir Kim, Bryan Russell, Mathieu Aubry
We propose to represent shapes as the deformation and combination of learnt elementary 3D structures. We demonstrate this decomposition in learnt elementary 3D structures is highly interpretable and leads to clear improvements in 3D shape generation and matching.

More precisely, we present two complementary approaches to learn elementary structures in a deep learning framework: (i) continuous surface deformation learning and (ii) 3D structure points learning. Both approaches can be extended to abstract structures of higher dimensions for improved results. We evaluate our method on two very different tasks: ShapeNet objects reconstruction and dense correspondences estimation between human scans. Qualitatively our approach provides interpretable and repeatable results. Quantitatively, we show an important 16% boost for 3D object generation via surface deformation, as well as a clear 6% improvement over state of the art correspondence results on the FAUST inter challenge.
****

Cross-Modal Learning with Adversarial Samples
CHAO LI, Shangqian Gao, Cheng Deng, De Xie, Wei Liu
With the rapid developments of deep neural networks, numerous deep cross-modal analysis methods have been presented and are being applied in widespread real-world applications, including healthcare and safety-critical environments. However, the recent studies on robustness and stability of deep neural networks show that a microscopic modification, known as adversarial sample, which is even imperceptible to humans, can easily fool a well-performed deep neural network and brings a new obstacle to deep cross-modal correlation exploring. In this paper, we propose a novel Cross-Modal correlation Learning with Adversarial samples, namely CMLA, which for the first time presents the existence of adversarial samples in cross-modal data. Moreover, we provide a simple yet effective adversarial sample learning method, where inter- and intra- modality similarity regularizations across different modalities are simultaneously integrated into the learning of adversarial samples. Finally, our proposed CMLA is demonstrated to be highly effective in cross-modal hashing based retrieval. Extensive experiments on two cross-modal benchmark datasets show that the adversarial examples produced by our CMLA are efficient in fooling a target deep cross-modal hashing network. On the other hand, such adversarial examples can significantly strengthen the robustness of the target network by conducting an adversarial training.
****

Learning Disentangled Representation for Robust Person Re-identification
Chanho Eom, Bumsub Ham
We address the problem of person re-identification (reID), that is, retrieving person images from a large dataset, given a query image of the person of interest. The key challenge is to learn person representations robust to intra-class variations, as different persons can have the same attribute and the same person's appearance looks different with viewpoint changes. Recent reID methods focus on learning discriminative features but robust to only a particular factor of variations (e.g., human pose) and this requires corresponding supervisory signals (e.g., pose annotations). To tackle this problem, we propose to disentangle identity-related and -unrelated features from person images. Identity-related features contain information useful for specifying a particular person (e.g.,clothing), while identity-unrelated ones hold other factors (e.g., human pose, scale changes). To this end, we introduce a new generative adversarial network, dubbed identity shuffle GAN (IS-GAN), that factorizes these features using identification labels without any auxiliary information. We also propose an identity shuffling technique to regularize the disentangled features. Experimental results demonstrate the effectiveness of IS-GAN, largely outperforming the state of the art on standard reID benchmarks including the Market-1501, CUHK03 and DukeMTMC-reID. Our code and models will be available online at the time of the publication.
****

On Testing for Biases in Peer Review
Ivan Stelmakh, Nihar Shah, Aarti Singh
We consider the issue of biases in scholarly research, specifically, in peer review. There is a long standing debate on whether exposing author identities to reviewers induces biases against certain groups, and our focus is on designing tests to detect the presence of such biases. Our starting point is a remarkable recent work by Tomkins, Zhang and Heavlin which conducted a controlled, large-scale experiment to investigate existence of biases in the peer reviewing of the WSDM

conference. We present two sets of results in this paper. The first set of results is negative, and pertains to the statistical tests and the experimental setup used in the work of Tomkins et al. We show that the test employed therein does not guarantee control over false alarm probability and under correlations between relevant variables, coupled with any of the following conditions, with high probability can declare a presence of bias when it is in fact absent: (a) measurement error, (b) model mismatch, (c) reviewer calibration. Moreover, we show that the setup of their experiment may itself inflate false alarm probability if (d) bidding is performed in non-blind manner or (e) popular reviewer assignment procedure is employed. Our second set of results is positive, in that we present a general framework for testing for biases in (single vs. double blind) peer review. We then present a hypothesis test with guaranteed control over false alarm probability and non-trivial power even under conditions (a)--(c). Conditions (d) and (e) are more fundamental problems that are tied to the experimental setup and not necessarily related to the test.
****
Learning Deterministic Weighted Automata with Queries and Counterexamples
Gail Weiss, Yoav Goldberg, Eran Yahav
We present an algorithm for reconstruction of a probabilistic deterministic finite automaton (PDFA) from a given black-box language model, such as a recurrent neural network (RNN).
The algorithm is a variant of the exact-learning algorithm L*, adapted to work in a probabilistic setting under noise.
The key insight of the adaptation is the use of conditional probabilities when making observations on the model, and the introduction of a variation tolerance when comparing observations.
When applied to RNNs, our algorithm returns models with better or equal word error rate (WER) and normalised distributed cumulative gain (NDCG) than achieved by n-gram or weighted finite automata (WFA) approximations of the same networks. The PDFAs capture a richer class of languages than n-grams, and are guaranteed to be stochastic and deterministic -- unlike the WFAs.
****
Making the Cut: A Bandit-based Approach to Tiered Interviewing
Candice Schumann, Zhi Lang, Jeffrey Foster, John Dickerson
Given a huge set of applicants, how should a firm allocate sequential resume screenings, phone interviews, and in-person site visits? In a tiered interview process, later stages (e.g., in-person visits) are more informative, but also more expensive than earlier stages (e.g., resume screenings). Using accepted hiring models and the concept of structured interviews, a best practice in human resources, we cast tiered hiring as a combinatorial pure exploration (CPE) problem in the stochastic multi-armed bandit setting. The goal is to select a subset of arms (in our case, applicants) with some combinatorial structure. We present new algorithms in both the probably approximately correct (PAC) and fixed-budget settings that select a near-optimal cohort with provable guarantees. We show via simulations on real data from one of the largest US-based computer science graduate programs that our algorithms make better hiring decisions or use less budget than the status quo.
****
Manifold-regression to predict from MEG/EEG brain signals without source modeling
David Sabbagh, Pierre Ablin, Gael Varoquaux, Alexandre Gramfort, Denis A. Engemann
Magnetoencephalography and electroencephalography (M/EEG) can reveal neuronal dynamics non-invasively in real-time and are therefore appreciated methods in medicine and neuroscience. Recent advances in modeling brain-behavior relationships have highlighted the effectiveness of Riemannian geometry for summarizing the spatially correlated time-series from M/EEG in terms of their covariance. However, after artefact-suppression, M/EEG data is often rank deficient which limits the application of Riemannian concepts. In this article, we focus on the task of regression with rank-reduced covariance matrices. We study two Riemannian approach

es that vectorize the M/EEG covariance between sensors through projection into a tangent space. The Wasserstein distance readily applies to rank-reduced data but lacks affine-invariance. This can be overcome by finding a common subspace in which the covariance matrices are full rank, enabling the affine-invariant geometric distance. We investigated the implications of these two approaches in synthetic generative models, which allowed us to control estimation bias of a linear model for prediction. We show that Wasserstein and geometric distances allow perfect out-of-sample prediction on the generative models. We then evaluated the methods on real data with regard to their effectiveness in predicting age from M/EEG covariance matrices. The findings suggest that the data-driven Riemannian methods outperform different sensor-space estimators and that they get close to the performance of biophysics-driven source-localization model that requires MRI acquisitions and tedious data processing. Our study suggests that the proposed Riemannian methods can serve as fundamental building-blocks for automated large-scale analysis of M/EEG.
****
Reflection Separation using a Pair of Unpolarized and Polarized Images
Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, Boxin Shi
When we take photos through glass windows or doors, the transmitted background scene is often blended with undesirable reflection. Separating two layers apart to enhance the image quality is of vital importance for both human and machine perception. In this paper, we propose to exploit physical constraints from a pair of unpolarized and polarized images to separate reflection and transmission layers. Due to the simplified capturing setup, the system becomes more underdetermined compared with existing polarization based solutions that take three or more images as input. We propose to solve semireflector orientation estimation first to make the physical image formation well-posed and then learn to reliably separate two layers using a refinement network with gradient loss. Quantitative and qualitative experimental results show our approach performs favorably over existing polarization and single image based solutions.
****
Co-Generation with GANs using AIS based HMC
Tiantian Fang, Alexander Schwing
Inferring the most likely configuration for a subset of variables of a joint distribution given the remaining ones -- which we refer to as co-generation -- is an important challenge that is computationally demanding for all but the simplest settings. This task has received a considerable amount of attention, particularly for classical ways of modeling distributions like structured prediction. In contrast, almost nothing is known about this task when considering recently proposed techniques for modeling high-dimensional distributions, particularly generative adversarial nets (GANs). Therefore, in this paper, we study the occurring challenges for co-generation with GANs. To address those challenges we develop an annealed importance sampling based Hamiltonian Monte Carlo co-generation algorithm. The presented approach significantly outperforms classical gradient based methods on a synthetic and on the CelebA and LSUN datasets.
****
Sim2real transfer learning for 3D human pose estimation: motion to the rescue
Carl Doersch, Andrew Zisserman
Synthetic visual data can provide practicically infinite diversity and rich labels,
while avoiding ethical issues with privacy and bias. However, for many tasks,
current models trained on synthetic data generalize poorly to real data. The task of
3D human pose estimation is a particularly interesting example of this sim2real
problem, because learning-based approaches perform reasonably well given real
training data, yet labeled 3D poses are extremely difficult to obtain in the wild,
limiting scalability. In this paper, we show that standard neural-network approaches,
which perform poorly when trained on synthetic RGB images, can perform well

when the data is pre-processed to extract cues about the person’s motion, notabl
y
as optical flow and the motion of 2D keypoints. Therefore, our results suggest
that motion can be a simple way to bridge a sim2real gap when video is available
.
We evaluate on the 3D Poses in the Wild dataset, the most challenging modern
benchmark for 3D pose estimation, where we show full 3D mesh recovery that is
on par with state-of-the-art methods trained on real 3D sequences, despite train
ing
only on synthetic humans from the SURREAL dataset.
****
Dimension-Free Bounds for Low-Precision Training
Zheng Li, Christopher M. De Sa
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 Disparate Impact of Personalized Interventions: Identifiability and Bounds
Nathan Kallus, Angela Zhou
Personalized interventions in social services, education, and healthcare leverage individual-level causal effect predictions in order to give the best treatment to each individual or to prioritize program interventions for the individuals most likely to benefit. While the sensitivity of these domains compels us to evaluate the fairness of such policies, we show that actually auditing their disparate impacts per standard observational metrics, such as true positive rates, is impossible since ground truths are unknown. Whether our data is experimental or observational, an individual's actual outcome under an intervention different than that received can never be known, only predicted based on features. We prove how we can nonetheless point-identify these quantities under the additional assumption of monotone treatment response, which may be reasonable in many applications. We further provide a sensitivity analysis for this assumption via sharp partial-identification bounds under violations of monotonicity of varying strengths. We show how to use our results to audit personalized interventions using partially-identified ROC and xROC curves and demonstrate this in a case study of a French job training dataset.
****
Cascade RPN: Delving into High-Quality Region Proposal Network with Adaptive Convolution
Thang Vu, Hyunjun Jang, Trung X. Pham, Chang Yoo
This paper considers an architecture referred to as Cascade Region Proposal Network (Cascade RPN) for improving the region-proposal quality and detection performance by systematically addressing the limitation of the conventional RPN that heuristically defines the anchors and aligns the features to the anchors. First, instead of using multiple anchors with predefined scales and aspect ratios, Cascade RPN relies on a single anchor per location and performs multi-stage refinement. Each stage is progressively more stringent in defining positive samples by starting out with an anchor-free metric followed by anchor-based metrics in the ensuing stages. Second, to attain alignment between the features and the anchors throughout the stages, adaptive convolution is proposed that takes the anchors in addition to the image features as its input and learns the sampled features guided by the anchors. A simple implementation of a two-stage Cascade RPN achieves 13.4 point AR higher than that of the conventional RPN, surpassing any existing region proposal methods. When adopting to Fast R-CNN and Faster R-CNN, Cascade RPN can improve the detection mAP by 3.1 and 3.5 points, respectively. The code will be made publicly available at https://github.com/thangvubk/Cascade-RPN.
****
Variational Bayesian Optimal Experimental Design
Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thoma

s Rainforth, Noah Goodman
Bayesian optimal experimental design (BOED) is a principled framework for making efficient use of limited experimental resources. Unfortunately, its applicability is hampered by the difficulty of obtaining accurate estimates of the expected information gain (EIG) of an experiment. To address this, we introduce several classes of fast EIG estimators by building on ideas from amortized variational inference. We show theoretically and empirically that these estimators can provide significant gains in speed and accuracy over previous approaches. We further demonstrate the practicality of our approach on a number of end-to-end experiments.
****
Flexible Modeling of Diversity with Strongly Log-Concave Distributions
Joshua Robinson, Suvrit Sra, Stefanie Jegelka
Strongly log-concave (SLC) distributions are a rich class of discrete probability distributions over subsets of some ground set. They are strictly more general than strongly Rayleigh (SR) distributions such as the well-known determinantal point process. While SR distributions offer elegant models of diversity, they lack an easy control over how they express diversity. We propose SLC as the right extension of SR that enables easier, more intuitive control over diversity, illustrating this via examples of practical importance. We develop two fundamental tools needed to apply SLC distributions to learning and inference: sampling and mode finding. For sampling we develop an MCMC sampler and give theoretical mixing time bounds. For mode finding, we establish a weak log-submodularity property for SLC functions and derive optimization guarantees for a distorted greedy algorithm.
****
Neural Machine Translation with Soft Prototype
Yiren Wang, Yingce Xia, Fei Tian, Fei Gao, Tao Qin, Cheng Xiang Zhai, Tie-Yan Liu
Neural machine translation models usually use the encoder-decoder framework and generate translation from left to right (or right to left) without fully utilizing the target-side global information. A few recent approaches seek to exploit the global information through two-pass decoding, yet have limitations in translation quality and model efficiency. In this work, we propose a new framework that introduces a soft prototype into the encoder-decoder architecture, which allows the decoder to have indirect access to both past and future information, such that each target word can be generated based on the better global understanding. We further provide an efficient and effective method to generate the prototype. Empirical studies on various neural machine translation tasks show that our approach brings significant improvement in generation quality over the baseline model, with little extra cost in storage and inference time, demonstrating the effectiveness of our proposed framework. Specially, we achieve state-of-the-art results on WMT2014, 2015 and 2017 English to German translation.
****
Unsupervised Curricula for Visual Meta-Reinforcement Learning
Allan Jabri, Kyle Hsu, Abhishek Gupta, Ben Eysenbach, Sergey Levine, Chelsea Finn
In principle, meta-reinforcement learning algorithms leverage experience across many tasks to learn fast and effective reinforcement learning (RL) strategies. However, current meta-RL approaches rely on manually-defined distributions of training tasks, and hand-crafting these task distributions can be challenging and time-consuming. Can ``useful'' pre-training tasks be discovered in an unsupervised manner? We develop an unsupervised algorithm for inducing an adaptive meta-training task distribution, i.e. an automatic curriculum, by modeling unsupervised interaction in a visual environment.
The task distribution is scaffolded by a parametric density model of the meta-learner's trajectory distribution.
We formulate unsupervised meta-RL as information maximization between a latent task variable and the meta-learner’s data distribution, and describe a practical instantiation which alternates between integration of recent experience into the

 task distribution and meta-learning of the updated tasks. Repeating this proced
ure leads to iterative reorganization such that the curriculum adapts as the met
a-learner's data distribution shifts. Moreover, we show how discriminative clust
ering frameworks for visual representations can support trajectory-level task a
cquisition and exploration in domains with pixel observations, avoiding the pitf
alls of alternatives.
In experiments on vision-based navigation and manipulation domains, we show that
 the algorithm allows for unsupervised meta-learning that both transfers to down
stream tasks specified by hand-crafted reward functions and serves as pre-traini
ng for more efficient meta-learning of test task distributions.
****
Improved Regret Bounds for Bandit Combinatorial Optimization
Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori
 Kakimura, Ken-Ichi Kawarabayashi
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.
****
Doubly-Robust Lasso Bandit
Gi-Soo Kim, Myunghee Cho Paik
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.
****
Recurrent Kernel Networks
Dexiong Chen, Laurent Jacob, Julien Mairal
Substring kernels are classical tools for representing biological sequences or t
ext. However, when large amounts of annotated data is available, models that all
ow end-to-end training such as neural networks are often prefered. Links betwee
n recurrent neural networks (RNNs) and substring kernels have recently been draw
n, by formally showing that RNNs with specific activation functions were points
in a reproducing kernel Hilbert space (RKHS). In this paper, we revisit this li
nk by generalizing convolutional kernel networks---originally related to a relax
ation of the mismatch kernel---to model gaps in sequences. It results in a new t
ype of recurrent neural network which can be trained end-to-end with backpropaga
tion, or without supervision by using kernel approximation techniques. We exper
imentally show that our approach is well suited to biological sequences, where i
t outperforms existing methods for protein classification tasks.
****
Thinning for Accelerating the Learning of Point Processes
Tianbo Li, Yiping Ke
This paper discusses one of the most fundamental issues about point processes th
at what is the best sampling method for point processes. We propose \textit{thin
ning} as a downsampling method for accelerating the learning of point processes.
 We find that the thinning operation preserves the structure of intensity, and i
s able to estimate parameters with less time and without much loss of accuracy.
Theoretical results including intensity, parameter and gradient estimation on a
thinned history are presented for point processes with decouplable intensities.
A stochastic optimization algorithm based on the thinned gradient is proposed. E
xperimental results on synthetic and real-world datasets validate the effectiven
ess of thinning in the tasks of parameter and gradient estimation, as well as st
ochastic optimization.
****
A Universally Optimal Multistage Accelerated Stochastic Gradient Method
Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, Asuman Ozdaglar
We study the problem of minimizing a strongly convex, smooth function when we ha
ve noisy estimates of its gradient. We propose a novel multistage accelerated al
gorithm that is universally optimal in the sense that it achieves the optimal ra

te both in the deterministic and stochastic case and operates without knowledge of noise characteristics. The algorithm consists of stages that use a stochastic version of Nesterov's method with a specific restart and parameters selected to achieve the fastest reduction in the bias-variance terms in the convergence rate bounds.
****
Ask not what AI can do, but what AI should do: Towards a framework of task delegability
Brian Lubars, Chenhao Tan
While artificial intelligence (AI) holds promise for addressing societal challenges, issues of exactly which tasks to automate and to what extent to do so remain understudied. We approach this problem of task delegability from a human-centered perspective by developing a framework on human perception of task delegation to AI. We consider four high-level factors that can contribute to a delegation decision: motivation, difficulty, risk, and trust. To obtain an empirical understanding of human preferences in different tasks, we build a dataset of 100 tasks from academic papers, popular media portrayal of AI, and everyday life, and administer a survey based on our proposed framework. We find little preference for full AI control and a strong preference for machine-in-the-loop designs, in which humans play the leading role. Among the four factors, trust is the most correlated with human preferences of optimal human-machine delegation. This framework represents a first step towards characterizing human preferences of AI automation across tasks. We hope this work encourages future efforts towards understanding such individual attitudes; our goal is to inform the public and the AI research community rather than dictating any direction in technology development.
****
Offline Contextual Bandits with High Probability Fairness Guarantees
Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, Philip S. Thomas
We present RobinHood, an of■ine contextual bandit algorithm designed to satisfy a broad family of fairness constraints. Our algorithm accepts multiple fairness de■nitions and allows users to construct their own unique fairness de■nitions for the problem at hand. We provide a theoretical analysis of RobinHood, which includes a proof that it will not return an unfair solution with probability greater than a user-speci■ed threshold. We validate our algorithm on three applications: a tutoring system in which we conduct a user study and consider multiple unique fairness de■nitions; a loan approval setting (using the Statlog German credit data set) in which well-known fairness de■nitions are applied; and criminal recidivism (using data released by ProPublica). In each setting, our algorithm is able to produce fair policies that achieve performance competitive with other of■ine and online contextual bandit algorithms.
****
Bias Correction of Learned Generative Models using Likelihood-Free Importance Weighting
Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J. Horvitz, Stefano Ermon
A learned generative model often produces biased statistics relative to the underlying data distribution. A standard technique to correct this bias is importance sampling, where samples from the model are weighted by the likelihood ratio under model and true distributions. When the likelihood ratio is unknown, it can be estimated by training a probabilistic classifier to distinguish samples from the two distributions. We employ this likelihood-free importance weighting method to correct for the bias in generative models. We find that this technique consistently improves standard goodness-of-fit metrics for evaluating the sample quality of state-of-the-art deep generative models, suggesting reduced bias. Finally, we demonstrate its utility on representative applications in a) data augmentation for classification using generative adversarial networks, and b) model-based policy evaluation using off-policy data.
****
LCA: Loss Change Allocation for Neural Network Training

Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski
Neural networks enjoy widespread use, but many aspects of their training, repres entation, and operation are poorly understood. In particular, our view into the training process is limited, with a single scalar loss being the most common vie wport into this high-dimensional, dynamic process. We propose a new window into training called Loss Change Allocation (LCA), in which credit for changes to the network loss is conservatively partitioned to the parameters. This measurement is accomplished by decomposing the components of an approximate path integral al ong the training trajectory using a Runge-Kutta integrator. This rich view shows which parameters are responsible for decreasing or increasing the loss during t raining, or which parameters "help" or "hurt" the network's learning, respective ly. LCA may be summed over training iterations and/or over neurons, channels, or layers for increasingly coarse views. This new measurement device produces seve ral insights into training. (1) We find that barely over 50% of parameters help during any given iteration. (2) Some entire layers hurt overall, moving on avera ge against the training gradient, a phenomenon we hypothesize may be due to phas e lag in an oscillatory training process. (3) Finally, increments in learning pr oceed in a synchronized manner across layers, often peaking on identical iterati ons.
****
Adaptive Cross-Modal Few-shot Learning
Chen Xing, Negar Rostamzadeh, Boris Oreshkin, Pedro O. O. Pinheiro
Metric-based meta-learning techniques have successfully been applied to few-shot classification problems. In this paper, we propose to leverage cross-modal info rmation to enhance metric-based few-shot learning methods.
Visual and semantic feature spaces have different structures by definition. For certain concepts, visual features might be richer and more discriminative than t ext ones. While for others, the inverse might be true. Moreover, when the suppor t from visual information is limited in image classification, semantic represent ations (learned from unsupervised text corpora) can provide strong prior knowled ge and context to help learning. Based on these two intuitions, we propose a mec hanism that can adaptively combine information from both modalities according to new image categories to be learned. Through a series of experiments, we show th at by this adaptive combination of the two modalities, our model outperforms cur rent uni-modality few-shot learning methods and modality-alignment methods by a large margin on all benchmarks and few-shot scenarios tested. Experiments also s how that our model can effectively adjust its focus on the two modalities. The improvement in performance is particularly large when the number of shots is ver y small.
****
Polynomial Cost of Adaptation for X-Armed Bandits
Hedi Hadiji
In the context of stochastic continuum-armed bandits, we present an algorithm th at adapts to the unknown smoothness of the objective function. We exhibit and co mpute a polynomial cost of adaptation to the Hölder regularity for regret minimi zation. To do this, we first reconsider the recent lower bound of Locatelli and Carpentier, 2018, and define and characterize admissible rate functions. Our new algorithm matches any of these minimal rate functions. We provide a finite-time analysis and a thorough discussion about asymptotic optimality.
****
Modelling heterogeneous distributions with an Uncountable Mixture of Asymmetric Laplacians
Axel Brando, Jose A. Rodriguez, Jordi Vitria, Alberto Rubio Muñoz
In regression tasks, aleatoric uncertainty is commonly addressed by considering a parametric distribution of the output variable, which is based on strong assum ptions such as symmetry, unimodality or by supposing a restricted shape. These a ssumptions are too limited in scenarios where complex shapes, strong skews or mu ltiple modes are present. In this paper, we propose a generic deep learning fram ework that learns an Uncountable Mixture of Asymmetric Laplacians (UMAL), which will allow us to estimate heterogeneous distributions of the output variable and

shows its connections to quantile regression. Despite having a fixed number of parameters, the model can be interpreted as an infinite mixture of components, which yields a flexible approximation for heterogeneous distributions. Apart from synthetic cases, we apply this model to room price forecasting and to predict financial operations in personal bank accounts. We demonstrate that UMAL produces proper distributions, which allows us to extract richer insights and to sharpen decision-making.
****

GNNExplainer: Generating Explanations for Graph Neural Networks
Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, Jure Leskovec
Graph Neural Networks (GNNs) are a powerful tool for machine learning on graphs. GNNs combine node feature information with the graph structure by recursively passing neural messages along edges of the input graph. However, incorporating both graph structure and feature information leads to complex models, and explaining predictions made by GNNs remains unsolved. Here we propose GNNExplainer, the first general, model-agnostic approach for providing interpretable explanations for predictions of any GNN-based model on any graph-based machine learning task. Given an instance, GNNExplainer identifies a compact subgraph structure and a small subset of node features that have a crucial role in GNN's prediction.
Further, GNNExplainer can generate consistent and concise explanations for an entire class of instances.
We formulate GNNExplainer as an optimization task that maximizes the mutual information between a GNN's prediction and distribution of possible subgraph structures. Experiments on synthetic and real-world graphs show that our approach can identify important graph structures as well as node features, and outperforms baselines by 17.1% on average. GNNExplainer provides a variety of benefits, from the ability to visualize semantically relevant structures to interpretability, to giving insights into errors of faulty GNNs.
****

Missing Not at Random in Matrix Completion: The Effectiveness of Estimating Missingness Probabilities Under a Low Nuclear Norm Assumption
Wei Ma, George H. Chen
Matrix completion is often applied to data with entries missing not at random (MNAR). For example, consider a recommendation system where users tend to only reveal ratings for items they like. In this case, a matrix completion method that relies on entries being revealed at uniformly sampled row and column indices can yield overly optimistic predictions of unseen user ratings. Recently, various papers have shown that we can reduce this bias in MNAR matrix completion if we know the probabilities of different matrix entries being missing. These probabilities are typically modeled using logistic regression or naive Bayes, which make strong assumptions and lack guarantees on the accuracy of the estimated probabilities. In this paper, we suggest a simple approach to estimating these probabilities that avoids these shortcomings. Our approach follows from the observation that missingness patterns in real data often exhibit low nuclear norm structure. We can then estimate the missingness probabilities by feeding the (always fully-observed) binary matrix specifying which entries are revealed to an existing nuclear-norm-constrained matrix completion algorithm by Davenport et al. [2014]. Thus, we tackle MNAR matrix completion by solving a different matrix completion problem first that recovers missingness probabilities. We establish finite-sample error bounds for how accurate these probability estimates are and how well these estimates debias standard matrix completion losses for the original matrix to be completed. Our experiments show that the proposed debiasing strategy can improve a variety of existing matrix completion algorithms, and achieves downstream matrix completion accuracy at least as good as logistic regression and naive Bayes debiasing baselines that require additional auxiliary information.
****

Unsupervised learning of object structure and dynamics from videos
Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P. Murphy, Honglak Lee

Extracting and predicting object structure and dynamics from videos without supervision is a major challenge in machine learning. To address this challenge, we adopt a keypoint-based image representation and learn a stochastic dynamics model of the keypoints. Future frames are reconstructed from the keypoints and a reference frame. By modeling dynamics in the keypoint coordinate space, we achieve stable learning and avoid compounding of errors in pixel space. Our method improves upon unstructured representations both for pixel-level video prediction and for downstream tasks requiring object-level understanding of motion dynamics. We evaluate our model on diverse datasets: a multi-agent sports dataset, the Human3.6M dataset, and datasets based on continuous control tasks from the DeepMind Control Suite. The spatially structured representation outperforms unstructured representations on a range of motion-related tasks such as object tracking, action recognition and reward prediction.
****
Scalable Structure Learning of Continuous-Time Bayesian Networks from Incomplete Data
Dominik Linzner, Michael Schmidt, Heinz Koeppl
Continuous-time Bayesian Networks (CTBNs) represent a compact yet powerful framework for understanding multivariate time-series data. Given complete data, parameters and structure can be estimated efficiently in closed-form. However, if data is incomplete, the latent states of the CTBN have to be estimated by laboriously simulating the intractable dynamics of the assumed CTBN. This is a problem, especially for structure learning tasks, where this has to be done for each element of a super-exponentially growing set of possible structures. In order to circumvent this notorious bottleneck, we develop a novel gradient-based approach to structure learning. Instead of sampling and scoring all possible structures individually, we assume the generator of the CTBN to be composed as a mixture of generators stemming from different structures. In this framework, structure learning can be performed via a gradient-based optimization of mixture weights. We combine this approach with a new variational method that allows for a closed-form calculation of this mixture marginal likelihood.
We show the scalability of our method by learning structures of previously inaccessible sizes from synthetic and real-world data.
****
Cross-channel Communication Networks
Jianwei Yang, Zhile Ren, Chuang Gan, Hongyuan Zhu, Devi Parikh
Convolutional neural networks process input data by sending channel-wise feature response maps to subsequent layers. While a lot of progress has been made by making networks deeper, information from each channel can only be propagated from lower levels to higher levels in a hierarchical feed-forward manner. When viewing each filter in the convolutional layer as a neuron, those neurons are not communicating explicitly within each layer in CNNs. We introduce a novel network unit called Cross-channel Communication (C3) block, a simple yet effective module to encourage the neuron communication within the same layer. The C3 block enables neurons to exchange information through a micro neural network, which consists of a feature encoder, a message communicator, and a feature decoder, before sending the information to the next layer. With C3 block, each neuron accounts for the channel-wise responses from other neurons at the same layer and learns more discriminative and complementary representations. Extensive experiments for multiple computer vision tasks show that our proposed mechanism allows shallower networks to aggregate useful information within each layer, and performances outperform baseline deep networks and other competitive methods.
****
Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training
Haichao Zhang, Jianyu Wang
We introduce a feature scattering-based adversarial training approach for improving model robustness against adversarial attacks.
Conventional adversarial training approaches leverage a supervised scheme (either targeted or non-targeted) in generating attacks for training, which typically

suffer from issues such as label leaking as noted in recent works. Differently, the proposed approach generates adversarial images for training through feature scattering in the latent space, which is unsupervised in nature and avoids label leaking. More importantly, this new approach generates perturbed images in a collaborative fashion, taking the inter-sample relationships into consideration. We conduct analysis on model robustness and demonstrate the effectiveness of the proposed approach through extensively experiments on different datasets compared with state-of-the-art approaches.
****

Identifying Causal Effects via Context-specific Independence Relations
Santtu Tikka, Antti Hyttinen, Juha Karvanen
Causal effect identification considers whether an interventional probability distribution can be uniquely determined from a passively observed distribution in a given causal structure. If the generating system induces context-specific independence (CSI) relations, the existing identification procedures and criteria based on do-calculus are inherently incomplete. We show that deciding causal effect non-identifiability is NP-hard in the presence of CSIs. Motivated by this, we design a calculus and an automated search procedure for identifying causal effects in the presence of CSIs. The approach is provably sound and it includes standard do-calculus as a special case. With the approach we can obtain identifying formulas that were unobtainable previously, and demonstrate that a small number of CSI-relations may be sufficient to turn a previously non-identifiable instance to identifiable.
****

Differentiable Ranking and Sorting using Optimal Transport
Marco Cuturi, Olivier Teboul, Jean-Philippe Vert
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.
****

Ordered Memory
Yikang Shen, Shawn Tan, Arian Hosseini, Zhouhan Lin, Alessandro Sordoni, Aaron C. Courville
Stack-augmented recurrent neural networks (RNNs) have been of interest to the deep learning community for some time. However, the difficulty of training memory models remains a problem obstructing the widespread use of such models. In this paper, we propose the Ordered Memory architecture. Inspired by Ordered Neurons (Shen et al., 2018), we introduce a new attention-based mechanism and use its cumulative probability to control the writing and erasing operation of the memory. We also introduce a new Gated Recursive Cell to compose lower-level representations into higher-level representation. We demonstrate that our model achieves strong performance on the logical inference task (Bowman et al., 2015) and the ListOps (Nangia and Bowman, 2018) task. We can also interpret the model to retrieve the induced tree structure, and find that these induced structures align with the ground truth. Finally, we evaluate our model on the Stanford Sentiment Treebank tasks (Socher et al., 2013), and find that it performs comparatively with the state-of-the-art methods in the literature.
****

Approximating the Permanent by Sampling from Adaptive Partitions
Jonathan Kuck, Tri Dao, Hamid Rezatofighi, Ashish Sabharwal, Stefano Ermon
Computing the permanent of a non-negative matrix is a core problem with practical applications ranging from target tracking to statistical thermodynamics. However, this problem is also #P-complete, which leaves little hope for finding an exact solution that can be computed efficiently. While the problem admits a fully polynomial randomized approximation scheme, this method has seen little use because it is both inefficient in practice and difficult to implement. We present ADAPART, a simple and efficient method for exact sampling of permutations, each associated with a weight as determined by a matrix. ADAPART uses an adaptive, iterative partitioning strategy over permutations to convert any upper bounding m

ethod for the permanent into one that satisfies a desirable `nesting' property over the partition used. These samples are then used to construct tight bounds on the permanent which hold with a high probability. Empirically, ADAPART provides significant speedups (sometimes exceeding 50x) over prior work. We also empirically observe polynomial scaling in some cases. In the context of multi-target tracking, ADAPART allows us to use the optimal proposal distribution during particle filtering, leading to orders of magnitude fewer samples and improved tracking performance.
****
Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics
Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, David Sussillo
Recurrent neural networks (RNNs) are a widely used tool for modeling sequential data, yet they are often treated as inscrutable black boxes. Given a trained recurrent network, we would like to reverse engineer it--to obtain a quantitative, interpretable description of how it solves a particular task. Even for simple tasks, a detailed understanding of how recurrent networks work, or a prescription for how to develop such an understanding, remains elusive. In this work, we use tools from dynamical systems analysis to reverse engineer recurrent networks trained to perform sentiment classification, a foundational natural language processing task. Given a trained network, we find fixed points of the recurrent dynamics and linearize the nonlinear system around these fixed points. Despite their theoretical capacity to implement complex, high-dimensional computations, we find that trained networks converge to highly interpretable, low-dimensional representations. In particular, the topological structure of the fixed points and corresponding linearized dynamics reveal an approximate line attractor within the RNN, which we can use to quantitatively understand how the RNN solves the sentiment analysis task. Finally, we find this mechanism present across RNN architectures (including LSTMs, GRUs, and vanilla RNNs) trained on multiple datasets, suggesting that our findings are not unique to a particular architecture or dataset. Overall, these results demonstrate that surprisingly universal and human interpretable computations can arise across a range of recurrent networks.
****
Quaternion Knowledge Graph Embeddings
SHUAI ZHANG, Yi Tay, Lina Yao, Qi Liu
In this work, we move beyond the traditional complex-valued representations, introducing more expressive hypercomplex representations to model entities and relations for knowledge graph embeddings. More specifically, quaternion embeddings, hypercomplex-valued embeddings with three imaginary components, are utilized to represent entities. Relations are modelled as rotations in the quaternion space. The advantages of the proposed approach are: (1) Latent inter-dependencies (between all components) are aptly captured with Hamilton product, encouraging a more compact interaction between entities and relations; (2) Quaternions enable expressive rotation in four-dimensional space and have more degree of freedom than rotation in complex plane; (3) The proposed framework is a generalization of ComplEx on hypercomplex space while offering better geometrical interpretations, concurrently satisfying the key desiderata of relational representation learning (i.e., modeling symmetry, anti-symmetry and inversion). Experimental results demonstrate that our method achieves state-of-the-art performance on four well-established knowledge graph completion benchmarks.
****
Initialization of ReLUs for Dynamical Isometry
Rebekka Burkholz, Alina Dubatovka
Deep learning relies on good initialization schemes and hyperparameter choices prior to training a neural network. Random weight initializations induce random network ensembles, which give rise to the trainability, training speed, and sometimes also generalization ability of an instance. In addition, such ensembles provide theoretical insights into the space of candidate models of which one is selected during training. The results obtained so far rely on mean field approximat

ions that assume infinite layer width and that study average squared signals. We derive the joint signal output distribution exactly, without mean field assumptions, for fully-connected networks with Gaussian weights and biases, and analyze deviations from the mean field results. For rectified linear units, we further discuss limitations of the standard initialization scheme, such as its lack of dynamical isometry, and propose a simple alternative that overcomes these by initial parameter sharing.
****
On the Transfer of Inductive Bias from Simulation to the Real World: a New Disentanglement Dataset
Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, Stefan Bauer
Learning meaningful and compact representations with disentangled semantic aspects is considered to be of key importance in representation learning. Since real-world data is notoriously costly to collect, many recent state-of-the-art disentanglement models have heavily relied on synthetic toy data-sets. In this paper, we propose a novel data-set which consists of over 1 million images of physical 3D objects with seven factors of variation, such as object color, shape, size and position. In order to be able to control all the factors of variation precisely, we built an experimental platform where the objects are being moved by a robotic arm. In addition, we provide two more datasets which consist of simulations of the experimental setup. These datasets provide for the first time the possibility to systematically investigate how well different disentanglement methods perform on real data in comparison to simulation, and how simulated data can be leveraged to build better representations of the real world. We provide a first experimental study of these questions and our results indicate that learned models transfer poorly, but that model and hyperparameter selection is an effective means of transferring information to the real world.
****
Subquadratic High-Dimensional Hierarchical Clustering
Amir Abboud, Vincent Cohen-Addad, Hussein Houdrouge
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.
****
PowerSGD: Practical Low-Rank Gradient Compression for Distributed Optimization
Thijs Vogels, Sai Praneeth Karimireddy, Martin Jaggi
We study gradient compression methods to alleviate the communication bottleneck in data-parallel distributed optimization. Despite the significant attention received, current compression schemes either do not scale well, or fail to achieve the target test accuracy. We propose a low-rank gradient compressor that can i) compress gradients rapidly, ii) efficiently aggregate the compressed gradients using all-reduce, and iii) achieve test performance on par with SGD. The proposed algorithm is the only method evaluated that achieves consistent wall-clock speedups when benchmarked against regular SGD with an optimized communication backend. We demonstrate reduced training times for convolutional networks as well as LSTMs on common datasets.
****
Distribution oblivious, risk-aware algorithms for multi-armed   bandits with unbounded rewards
Anmol Kagrecha, Jayakrishnan Nair, Krishna Jagannathan
Classical multi-armed bandit problems use the expected value of an arm as a metric
to evaluate its goodness. However, the expected value is a risk-neutral metric. In
many applications like finance, one is interested in balancing the expected return
of an arm (or portfolio) with the risk associated with that return. In this pape

r,
we consider the problem of selecting the arm that optimizes a linear combination of the expected reward and the associated Conditional Value at Risk (CVaR) in a fixed budget best-arm identification framework. We allow the reward distribution s
to be unbounded or even heavy-tailed. For this problem, our goal is to devise algorithms that are entirely distribution oblivious, i.e., the algorithm is not aware of
any information on the reward distributions, including bounds on the moments/tai ls,
or the suboptimality gaps across arms.
In this paper, we provide a class of such algorithms with provable upper bounds on the probability of incorrect identification. In the process, we develop a nov el
estimator for the CVaR of unbounded (including heavy-tailed) random variables and prove a concentration inequality for the same, which could be of independent interest. We also compare the error bounds for our distribution oblivious algori thms
with those corresponding to standard non-oblivious algorithms. Finally, numerica l
experiments reveal that our algorithms perform competitively when compared with non-oblivious algorithms, suggesting that distribution obliviousness can be real ised
in practice without incurring a significant loss of performance.
****
Multilabel reductions: what is my loss optimising?
Aditya K. Menon, Ankit Singh Rawat, Sashank Reddi, Sanjiv Kumar
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.
****
A Similarity-preserving Network Trained on Transformed Images Recapitulates Sali ent Features of the Fly Motion Detection Circuit
Yanis Bahroun, Dmitri Chklovskii, Anirvan Sengupta
Learning to detect content-independent transformations from data is one of the c entral problems in biological and artificial intelligence. An example of such pr oblem is unsupervised learning of a visual motion detector from pairs of consecu tive video frames. Rao and Ruderman formulated this problem in terms of learning infinitesimal transformation operators (Lie group generators) via minimizing i mage reconstruction error. Unfortunately, it is difficult to map their model ont o a biologically plausible neural network (NN) with local learning rules. Here w e propose a biologically plausible model of motion detection. We also adopt the transformation-operator approach but, instead of reconstruction-error minimizati on, start with a similarity-preserving objective function. An online algorithm t hat optimizes such an objective function naturally maps onto an NN with biologic ally plausible learning rules. The trained NN recapitulates major features of th e well-studied motion detector in the fly. In particular, it is consistent with the experimental observation that local motion detectors combine information fro m at least three adjacent pixels, something that contradicts the celebrated Hass enstein-Reichardt model.
****
CNN^{2}: Viewpoint Generalization via a Binocular Vision
Wei-Da Chen, Shan-Hung (Brandon) Wu
The Convolutional Neural Networks (CNNs) have laid the foundation for many techn iques in various applications. Despite achieving remarkable performance in some tasks, the 3D viewpoint generalizability of CNNs is still far behind humans visu al capabilities. Although recent efforts, such as the Capsule Networks, have bee n made to address this issue, these new models are either hard to train and/or i ncompatible with existing CNN-based techniques specialized for different applica

tions. Observing that humans use binocular vision to understand the world, we st udy in this paper whether the 3D viewpoint generalizability of CNNs can be achie ved via a binocular vision. We propose CNN^{2}, a CNN that takes two images as i nput, which resembles the process of an object being viewed from the left eye an d the right eye. CNN^{2} uses novel augmentation, pooling, and convolutional lay ers to learn a sense of three-dimensionality in a recursive manner. Empirical ev aluation shows that CNN^{2} has improved viewpoint generalizability compared to vanilla CNNs. Furthermore, CNN^{2} is easy to implement and train, and is compat ible with existing CNN-based specialized techniques for different applications.
****
Unsupervised Learning of Object Keypoints for Perception and Control
Tejas D. Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Re ynolds, Andrew Zisserman, Volodymyr Mnih
The study of object representations in computer vision has primarily focused on developing representations that are useful for image classification, object dete ction, or semantic segmentation as downstream tasks. In this work we aim to lear n object representations that are useful for control and reinforcement learning (RL). To this end, we introduce Transporter, a neural network architecture for d iscovering concise geometric object representations in terms of keypoints or ima ge-space coordinates. Our method learns from raw video frames in a fully unsuper vised manner, by transporting learnt image features between video frames using a keypoint bottleneck. The discovered keypoints track objects and object parts ac ross long time-horizons more accurately than recent similar methods. Furthermore , consistent long-term tracking enables two notable results in control domains - - (1) using the keypoint co-ordinates and corresponding image features as inputs enables highly sample-efficient reinforcement learning; (2) learning to explore by controlling keypoint locations drastically reduces the search space, enablin g deep exploration (leading to states unreachable through random action explorat ion) without any extrinsic rewards.
****
G2SAT: Learning to Generate SAT Formulas
Jiaxuan You, Haoze Wu, Clark Barrett, Raghuram Ramanujan, Jure Leskovec
The Boolean Satisfiability (SAT) problem is the canonical NP-complete problem an d is fundamental to computer science, with a wide array of applications in plann ing, verification, and theorem proving. Developing and evaluating practical SAT solvers relies on extensive empirical testing on a set of real-world benchmark f ormulas. However, the availability of such real-world SAT formulas is limited. W hile these benchmark formulas can be augmented with synthetically generated ones , existing approaches for doing so are heavily hand-crafted and fail to simultan eously capture a wide range of characteristics exhibited by real-world SAT insta nces. In this work, we present G2SAT, the first deep generative framework that l earns to generate SAT formulas from a given set of input formulas. Our key insig ht is that SAT formulas can be transformed into latent bipartite graph represent ations which we model using a specialized deep generative neural network. We sho w that G2SAT can generate SAT formulas that closely resemble given real-world SA T instances, as measured by both graph metrics and SAT solver behavior. Further, we show that our synthetic SAT formulas could be used to improve SAT solver per formance on real-world benchmarks, which opens up new opportunities for the cont inued development of SAT solvers and a deeper understanding of their performance .
****
The Functional Neural Process
Christos Louizos, Xiahan Shi, Klamer Schutte, Max Welling
We present a new family of exchangeable stochastic processes, the Functional Neu ral Processes (FNPs). FNPs model distributions over functions by learning a grap h of dependencies on top of latent representations of the points in the given da taset. In doing so, they define a Bayesian model without explicitly positing a p rior distribution over latent global parameters; they instead adopt priors over the relational structure of the given dataset, a task that is much simpler. We s how how we can learn such models from data, demonstrate that they are scalable t

o large datasets through mini-batch optimization and describe how we can make predictions for new points via their posterior predictive distribution. We experimentally evaluate FNPs on the tasks of toy regression and image classification and show that, when compared to baselines that employ global latent parameters, they offer both competitive predictions as well as more robust uncertainty estimates.
****
Convergent Policy Optimization for Safe Reinforcement Learning
Ming Yu, Zhuoran Yang, Mladen Kolar, Zhaoran Wang
We study the safe reinforcement learning problem with nonlinear function approximation, where policy optimization is formulated as a constrained optimization problem with both the objective and the constraint being nonconvex functions. For such a problem, we construct a sequence of surrogate convex constrained optimization problems by replacing the nonconvex functions locally with convex quadratic functions obtained from policy gradient estimators. We prove that the solutions to these surrogate problems converge to a stationary point of the original nonconvex problem. Furthermore, to extend our theoretical results, we apply our algorithm to examples of optimal control and multi-agent reinforcement learning with safety constraints.
****
A Refined Margin Distribution Analysis for Forest Representation Learning
Shen-Huan Lyu, Liang Yang, 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.
****
Diffeomorphic Temporal Alignment Nets
Ron A. Shapira Weber, Matan Eyal, Nicki Skafte, Oren Shriki, Oren Freifeld
Time-series analysis is confounded by nonlinear time warping of the data. Traditional methods for joint alignment do not generalize: after aligning a given signal ensemble, they lack a mechanism, that does not require solving a new optimization problem, to align previously-unseen signals. In the multi-class case, they must also first classify the test data before aligning it. Here we propose the Diffeomorphic Temporal alignment Net (DTAN), a learning-based method for time-series joint alignment. Via flexible temporal transformer layers, DTAN learns and applies an input-dependent nonlinear time warping to its input signal. Once learned, DTAN easily aligns previously-unseen signals by its inexpensive forward pass. In a single-class case, the method is unsupervised: the ground-truth alignments are unknown. In the multi-class case, it is semi-supervised in the sense that class labels (but not the ground-truth alignments) are used during learning; in test time, however, the class labels are unknown. As we show, DTAN not only outperforms existing joint-alignment methods in aligning training data but also generalizes well to test data. Our code is available at https://github.com/BGU-CS-VIL/dtan.
****
Multi-source Domain Adaptation for Semantic Segmentation
Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, Kurt Keutzer
Simulation-to-real domain adaptation for semantic segmentation has been actively studied for various applications such as autonomous driving. Existing methods mainly focus on a single-source setting, which cannot easily handle a more practical scenario of multiple sources with different distributions. In this paper, we propose to investigate multi-source domain adaptation for semantic segmentation. Specifically, we design a novel framework, termed Multi-source Adversarial Domain Aggregation Network (MADAN), which can be trained in an end-to-end manner. First, we generate an adapted domain for each source with dynamic semantic consistency while aligning at the pixel-level cycle-consistently towards the target. Second, we propose sub-domain aggregation discriminator and cross-domain cycle discriminator to make different adapted domains more closely aggregated. Finally,

feature-level alignment is performed between the aggregated domain and target do main while training the segmentation network. Extensive experiments from synthet ic GTA and SYNTHIA to real Cityscapes and BDDS datasets demonstrate that the pro posed MADAN model outperforms state-of-the-art approaches. Our source code is r eleased at: https://github.com/Luodian/MADAN.
****
Spectral Modification of Graphs for Improved Spectral Clustering
Ioannis Koutis, Huong 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-auth ors prior to requesting a name change in the electronic proceedings.
****
On Exact Computation with an Infinitely Wide Neural Net
Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R. Salakhutdinov, Ruosong W ang
How well does a classic deep net architecture like AlexNet or VGG19 classify on a standard dataset such as CIFAR-10 when its “width”— namely, number of channels in convolutional layers, and number of nodes in fully-connected internal layers — is allowed to increase to infinity? Such questions have come to the forefront in the quest to theoretically understand deep learning and its mysteries about optimization and generalization. They also connect deep learning to notions such as Gaussian processes and kernels. A recent paper [Jacot et al., 2018] introduc ed the Neural Tangent Kernel (NTK) which captures the behavior of fully-connecte d deep nets in the infinite width limit trained by gradient descent; this object was implicit in some other recent papers. An attraction of such ideas is that a pure kernel-based method is used to capture the power of a fully-trained deep n et of infinite width.
****
Small ReLU networks are powerful memorizers: a tight analysis of memorization ca pacity
Chulhee Yun, 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-auth ors prior to requesting a name change in the electronic proceedings.
****
Amortized Bethe Free Energy Minimization for Learning MRFs
Sam Wiseman, Yoon Kim
We propose to learn deep undirected graphical models (i.e., MRFs) with a non-ELB O objective for which we can calculate exact gradients. In particular, we optimi ze a saddle-point objective deriving from the Bethe free energy approximation to the partition function. Unlike much recent work in approximate inference, the d erived objective requires no sampling, and can be efficiently computed even for very expressive MRFs. We furthermore amortize this optimization with trained inf erence networks. Experimentally, we find that the proposed approach compares fav orably with loopy belief propagation, but is faster, and it allows for attaining better held out log likelihood than other recent approximate inference schemes.
****
Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empiric al Evidence
Fengxiang He, Tongliang Liu, Dacheng Tao
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.
****
XLNet: Generalized Autoregressive Pretraining for Language Understanding
Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, Qu oc V. Le

With the capability of modeling bidirectional contexts, denoising autoencoding b
ased pretraining like BERT achieves better performance than pretraining approach
es based on autoregressive language modeling.
However, relying on corrupting the input with masks, BERT neglects dependency be
tween the masked positions and suffers from a pretrain-finetune discrepancy.
In light of these pros and cons, we propose XLNet, a generalized autoregressive
pretraining method that (1) enables learning bidirectional contexts by maximizin
g the expected likelihood over all permutations of the factorization order and (
2) overcomes the limitations of BERT thanks to its autoregressive formulation.
Furthermore, XLNet integrates ideas from Transformer-XL, the state-of-the-art au
toregressive model, into pretraining.
Empirically, under comparable experiment setting, XLNet outperforms BERT on 20 t
asks, often by a large margin, including question answering, natural language in
ference, sentiment analysis, and document ranking.
****
Conditional Independence Testing using Generative Adversarial Networks
Alexis Bellot, Mihaela van der Schaar
We consider the hypothesis testing problem of detecting conditional dependence,
with a focus on high-dimensional feature spaces. Our contribution is a new test
statistic based on samples from a generative adversarial network designed to app
roximate directly a conditional distribution that encodes the null hypothesis, i
n a manner that maximizes power (the rate of true negatives). We show that such
an approach requires only that density approximation be viable in order to ensur
e that we control type I error (the rate of false positives); in particular, no
assumptions need to be made on the form of the distributions or feature dependen
cies. Using synthetic simulations with high-dimensional data we demonstrate sign
ificant gains in power over competing methods. In addition, we illustrate the us
e of our test to discover causal markers of disease in genetic data.
****
A Tensorized Transformer for Language Modeling
Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, Dawei Son
g
Latest development of neural models has connected the encoder and decoder throug
h a self-attention mechanism. In particular, Transformer, which is solely based
on self-attention, has led to breakthroughs in Natural Language Processing (NLP)
 tasks. However, the multi-head attention mechanism, as a key component of Trans
former, limits the effective deployment of the model to a resource-limited setti
ng. In this paper, based on the ideas of tensor decomposition and parameters sha
ring, we propose a novel self-attention model (namely Multi-linear attention) wi
th Block-Term Tensor Decomposition (BTD). We test and verify the proposed attent
ion method on three language modeling tasks (i.e., PTB, WikiText-103 and One-bil
lion) and a neural machine translation task (i.e., WMT-2016 English-German). Mul
ti-linear attention can not only largely compress the model parameters but also
obtain performance improvements, compared with a number of language modeling app
roaches, such as Transformer, Transformer-XL, and Transformer with tensor train
decomposition.
****
Classification-by-Components: Probabilistic Modeling of Reasoning over a Set of
Components
Sascha Saralajew, Lars Holdijk, Maike Rees, Ebubekir Asan, Thomas Villmann
Abstract Neural networks are state-of-the-art classification approaches but are
generally difficult to interpret. This issue can be partly alleviated by constru
cting a precise decision process within the neural network. In this work, a netw
ork architecture, denoted as Classification-By-Components network (CBC), is prop
osed. It is restricted to follow an intuitive reasoning based decision process i
nspired by Biederman's recognition-by-components theory from cognitive psycholog
y. The network is trained to learn and detect generic components that characteri
ze objects. In parallel, a class-wise reasoning strategy based on these componen
ts is learned to solve the classification problem. In contrast to other work on
reasoning, we propose three different types of reasoning: positive, negative, an

d indefinite. These three types together form a probability space to provide a p robabilistic classifier. The decomposition of objects into generic components co mbined with the probabilistic reasoning provides by design a clear interpretatio n of the classification decision process. The evaluation of the approach on MNIS T shows that CBCs are viable classifiers. Additionally, we demonstrate that the inherent interpretability offers a profound understanding of the classification behavior such that we can explain the success of an adversarial attack. The meth od's scalability is successfully tested using the ImageNet dataset.
****

Recurrent Registration Neural Networks for Deformable Image Registration
Robin Sandkühler, Simon Andermatt, Grzegorz Bauman, Sylvia Nyilas, Christoph Jud , Philippe C. Cattin
Parametric spatial transformation models have been successfully applied to image registration tasks. In such models, the transformation of interest is parameteri zed
by a fixed set of basis functions as for example B-splines. Each basis function is located on a fixed regular grid position among the image domain because the transformation of interest is not known in advance. As a consequence, not all ba sis
functions will necessarily contribute to the final transformation which results in a
non-compact representation of the transformation. We reformulate the pairwise registration problem as a recursive sequence of successive alignments. For each element in the sequence, a local deformation defined by its position, shape, and weight is computed by our recurrent registration neural network. The sum of all lo-
cal deformations yield the final spatial alignment of both images. Formulating t he
registration problem in this way allows the network to detect non-aligned region s in
the images and to learn how to locally refine the registration properly. In cont rast to
current non-sequence-based registration methods, our approach iteratively applie s
local spatial deformations to the images until the desired registration accuracy is achieved. We trained our network on 2D magnetic resonance images of the lung and compared our method to a standard parametric B-spline registration. The experiments show, that our method performs on par for the accuracy but yields a more compact representation of the transformation. Furthermore, we achieve a speedup of around 15 compared to the B-spline registration.
****

User-Specified Local Differential Privacy in Unconstrained Adaptive Online Learn ing
Dirk van der Hoeven
Local differential privacy is a strong notion of privacy in which the provider o f the data guarantees privacy by perturbing the data with random noise. In the s tandard application of local differential differential privacy the distribution of the noise is constant and known by the learner. In this paper we generalize t his approach by allowing the provider of the data to choose the distribution of the noise without disclosing any parameters of the distribution to the learner, under the constraint that the distribution is symmetrical. We consider this prob lem in the unconstrained Online Convex Optimization setting with noisy feedback. In this setting the learner receives the subgradient of a loss function, pertur bed by noise, and aims to achieve sublinear regret with respect to some competit or, without constraints on the norm of the competitor. We derive the first algor ithms that have adaptive regret bounds in this setting, i.e. our algorithms adap t to the unknown competitor norm, unknown noise, and unknown sum of the norms of the subgradients, matching state of the art bounds in all cases.
****

Learning Representations by Maximizing Mutual Information Across Views

Philip Bachman, R Devon Hjelm, William Buchwalter
We propose an approach to self-supervised representation learning based on maxim izing mutual information between features extracted from multiple views of a sha red context. For example, one could produce multiple views of a local spatio-tem poral context by observing it from different locations (e.g., camera positions w ithin a scene), and via different modalities (e.g., tactile, auditory, or visual). Or, an ImageNet image could provide a context from which one produces multipl e views by repeatedly applying data augmentation. Maximizing mutual information between features extracted from these views requires capturing information about high-level factors whose influence spans multiple views – e.g., presence of cer tain objects or occurrence of certain events. Following our proposed approach, w e develop a model which learns image representations that significantly outperfo rm prior methods on the tasks we consider. Most notably, using self-supervised l earning, our model learns representations which achieve 68.1% accuracy on ImageN et using standard linear evaluation. This beats prior results by over 12% and c oncurrent results by 7%. When we extend our model to use mixture-based represent ations, segmentation behaviour emerges as a natural side-effect. Our code is ava ilable online: https://github.com/Philip-Bachman/amdim-public.
****
Exploration Bonus for Regret Minimization in Discrete and Continuous Average Rew ard MDPs
Jian QIAN, Ronan Fruit, Matteo Pirotta, 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-auth ors prior to requesting a name change in the electronic proceedings.
****
A neurally plausible model learns successor representations in partially observa ble environments
Eszter Vértes, Maneesh Sahani
Animals need to devise strategies to maximize returns while interacting with the ir environment based on incoming noisy sensory observations. Task-relevant state s, such as the agent's location within an environment or the presence of a preda tor, are often not directly observable but must be inferred using available sens ory information. Successor representations (SR) have been proposed as a middle-g round between model-based and model-free reinforcement learning strategies, allo wing for fast value computation and rapid adaptation to changes in the reward fu nction or goal locations. Indeed, recent studies suggest that features of neura l responses are consistent with the SR framework. However, it is not clear how such representations might be learned and computed in partially observed, noisy environments. Here, we introduce a neurally plausible model using \emph{distribu tional successor features}, which builds on the distributed distributional code for the representation and computation of uncertainty, and which allows for effi cient value function computation in partially observed environments via the succ essor representation. We show that distributional successor features can support reinforcement learning in noisy environments in which direct learning of succes sful policies is infeasible.
****
Tight Dimension Independent Lower Bound on the Expected Convergence Rate for Dim inishing Step Sizes in SGD
PHUONG_HA NGUYEN, Lam Nguyen, Marten van Dijk
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.
****
Cost Effective Active Search
Shali Jiang, Roman Garnett, Benjamin Moseley
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.
****
Optimal Statistical Rates for Decentralised Non-Parametric Regression with Linea r Speed-Up
Dominic Richards, Patrick Rebeschini
We analyse the learning performance of Distributed Gradient Descent in the conte xt of multi-agent decentralised non-parametric regression with the square loss f unction when i.i.d. samples are assigned to agents. We show that if agents hold sufficiently many samples with respect to the network size, then Distributed Gra dient Descent achieves optimal statistical rates with a number of iterations tha t scales, up to a threshold, with the inverse of the spectral gap of the gossip matrix divided by the number of samples owned by each agent raised to a problem-dependent power. The presence of the threshold comes from statistics. It encodes the existence of a "big data" regime where the number of required iterations do es not depend on the network topology. In this regime, Distributed Gradient Desc ent achieves optimal statistical rates with the same order of iterations as grad ient descent run with all the samples in the network. Provided the communication delay is sufficiently small, the distributed protocol yields a linear speed-up in runtime compared to the single-machine protocol. This is in contrast to decen tralised optimisation algorithms that do not exploit statistics and only yield a linear speed-up in graphs where the spectral gap is bounded away from zero. Our results exploit the statistical concentration of quantities held by agents and shed new light on the interplay between statistics and communication in decentra lised methods. Bounds are given in the standard non-parametric setting with sour ce/capacity assumptions.
****
Modeling Conceptual Understanding in Image Reference Games
Rodolfo Corona Rodriguez, Stephan Alaniz, Zeynep Akata
An agent who interacts with a wide population of other agents needs to be aware that there may be variations in their understanding of the world.
Furthermore, the machinery which they use to perceive may be inherently differen t, as is the case between humans and machines.
In this work, we present both an image reference game between a speaker and a po pulation of listeners where reasoning about the concepts other agents can compre hend is necessary and a model formulation with this capability.
We focus on reasoning about the conceptual understanding of others, as well as a dapting to novel gameplay partners and dealing with differences in perceptual ma chinery.
Our experiments on three benchmark image/attribute datasets suggest that our lea rner indeed encodes information directly pertaining to the understanding of othe r agents, and that leveraging this information is crucial for maximizing gamepla y performance.
****
Inherent Weight Normalization in Stochastic Neural Networks
Georgios Detorakis, Sourav Dutta, Abhishek Khanna, Matthew Jerry, Suman Datta, E mre Neftci
Multiplicative stochasticity such as Dropout improves the robustness and gener-alizability deep neural networks. Here, we further demonstrate that always-on multiplicative stochasticity combined with simple threshold neurons provide a su f-
ficient substrate for deep learning machines. We call such models Neural Samplin g Machines (NSM). We find that the probability of activation of the NSM exhibits a self-normalizing property that mirrors Weight Normalization, a previously stu died mechanism that fulfills many of the features of Batch Normalization in an o nline fashion. The normalization of activities during training speeds up converg ence by preventing internal covariate shift caused by changes in the distributio n of inputs. The always-on stochasticity of the NSM confers the following advant ages: the network is identical in the inference and learning phases, making the NSM a suitable substrate for continual learning, it can exploit stochasticity in

herent to a physical substrate such as analog non-volatile memories for in memory computing, and it is suitable for Monte Carlo sampling, while requiring almost exclusively addition and comparison operations. We demonstrate NSMs on standard classification benchmarks (MNIST and CIFAR) and event-based classification benchmarks (N-MNIST and DVS Gestures). Our results show that NSMs perform comparably or better than conventional artificial neural networks with the same architecture.
****
Universality in Learning from Linear Measurements
Ehsan Abbasi, Fariborz Salehi, 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.
****
Discrimination in Online Markets: Effects of Social Bias on Learning from Reviews and Policy Design
Faidra Georgia Monachou, Itai Ashlagi
The increasing popularity of online two-sided markets such as ride-sharing, accommodation and freelance labor platforms, goes hand in hand with new socioeconomic challenges. One major issue remains the existence of bias and discrimination against certain social groups. We study this problem using a two-sided large market model with employers and workers mediated by a platform. Employers who seek to hire workers face uncertainty about a candidate worker's skill level. Therefore, they base their hiring decision on learning from past reviews about an individual worker as well as on their (possibly misspecified) prior beliefs about the ability level of the social group the worker belongs to. Drawing upon the social learning literature with bounded rationality and limited information, uncertainty combined with social bias leads to unequal hiring opportunities between workers of different social groups. Although the effect of social bias decreases as the number of reviews increases (consistent with empirical findings), minority workers still receive lower expected payoffs. Finally, we consider a simple directed matching policy (DM), which combines learning and matching to make better matching decisions for minority workers. Under this policy, there exists a steady-state equilibrium, in which DM reduces the discrimination gap.
****
Structure Learning with Side Information: Sample Complexity
Saurabh Sihag, Ali Tajer
Graphical models encode the stochastic dependencies among random variables (RVs). The vertices represent the RVs, and the edges signify the conditional dependencies among the RVs. Structure learning is the process of inferring the edges by observing realizations of the RVs, and it has applications in a wide range of technological, social, and biological networks. Learning the structure of graphs when the vertices are treated in isolation from inferential information known about them is well-investigated. In a wide range of domains, however, often there exist additional inferred knowledge about the structure, which can serve as valuable side information. For instance, the gene networks that represent different subtypes of the same cancer share similar edges across all subtypes and also have exclusive edges corresponding to each subtype, rendering partially similar graphical models for gene expression in different cancer subtypes. Hence, an inferential decision regarding a gene network can serve as side information for inferring other related gene networks. When such side information is leveraged judiciously, it can translate to significant improvement in structure learning. Leveraging such side information can be abstracted as inferring structures of distinct graphical models that are {\sl partially} similar. This paper focuses on Ising graphical models, and considers the problem of simultaneously learning the structures of two {\sl partially} similar graphs, where any inference about the structure of one graph offers side information for the other graph. The bounded edge subclass of Ising models is considered, and necessary conditions (information-theoretic), as well as sufficient conditions (algorithmic) for the sample complex

ity for achieving a bounded probability of error, are established. Furthermore, specific regimes are identified in which the necessary and sufficient conditions coincide, rendering the optimal sample complexity.
****

Discrete Flows: Invertible Generative Models of Discrete Data
Dustin Tran, Keyon Vafa, Kumar Agrawal, Laurent Dinh, Ben Poole
While normalizing flows have led to significant advances in modeling high-dimensional continuous distributions, their applicability to discrete distributions remains unknown. In this paper, we show that flows can in fact be extended to discrete events---and under a simple change-of-variables formula not requiring log-determinant-Jacobian computations. Discrete flows have numerous applications. We consider two flow architectures: discrete autoregressive flows that enable bidirectionality, allowing, for example, tokens in text to depend on both left-to-right and right-to-left contexts in an exact language model; and discrete bipartite flows that enable efficient non-autoregressive generation as in RealNVP. Empirically, we find that discrete autoregressive flows outperform autoregressive baselines on synthetic discrete distributions, an addition task, and Potts models; and bipartite flows can obtain competitive performance with autoregressive baselines on character-level language modeling for Penn Tree Bank and text8.
****

Disentangled behavioural representations
Amir Dezfouli, Hassan Ashtiani, Omar Ghattas, Richard Nock, Peter Dayan, Cheng Soon Ong
Individual characteristics in human decision-making are often quantified by fitting a parametric cognitive model to subjects' behavior and then studying differences between them in the associated parameter space. However, these models often fit behavior more poorly than recurrent neural networks (RNNs), which are more flexible and make fewer assumptions about the underlying decision-making processes. Unfortunately, the parameter and latent activity spaces of RNNs are generally high-dimensional and uninterpretable, making it hard to use them to study individual differences. Here, we show how to benefit from the flexibility of RNNs while representing individual differences in a low-dimensional and interpretable space. To achieve this, we propose a novel end-to-end learning framework in which an encoder is trained to map the behavior of subjects into a low-dimensional latent space. These low-dimensional representations are used to generate the parameters of individual RNNs corresponding to the decision-making process of each subject. We introduce terms into the loss function that ensure that the latent dimensions are informative and disentangled, i.e., encouraged to have distinct effects on behavior. This allows them to align with separate facets of individual differences. We illustrate the performance of our framework on synthetic data as well as a dataset including the behavior of patients with psychiatric disorders.
****

A Flexible Generative Framework for Graph-based Semi-supervised Learning
Jiaqi Ma, Weijing Tang, Ji Zhu, Qiaozhu Mei
We consider a family of problems that are concerned about making predictions for the majority of unlabeled, graph-structured data samples based on a small proportion of labeled samples. Relational information among the data samples, often encoded in the graph/network structure, is shown to be helpful for these semi-supervised learning tasks. However, conventional graph-based regularization methods and recent graph neural networks do not fully leverage the interrelations between the features, the graph, and the labels. In this work, we propose a flexible generative framework for graph-based semi-supervised learning, which approaches the joint distribution of the node features, labels, and the graph structure. Borrowing insights from random graph models in network science literature, this j

oint distribution can be instantiated using various distribution families. For t
he inference of missing labels, we exploit recent advances of scalable variation
al inference techniques to approximate the Bayesian posterior. We conduct thorou
gh experiments on benchmark datasets for graph-based semi-supervised learning. R
esults show that the proposed methods outperform state-of-the-art models under m
ost settings.
****
A New Perspective on Pool-Based Active Classification and False-Discovery Contro
l
Lalit Jain, Kevin G. Jamieson
In many scientific settings there is a need for adaptive experimental design to
guide the process of identifying regions of the search space that contain as man
y true positives as possible subject to a low rate of false discoveries (i.e. fa
lse alarms). Such regions of the search space could differ drastically from a pr
edicted set that minimizes 0/1 error and accurate identification could require v
ery different sampling strategies. Like active learning for binary classificatio
n, this experimental design cannot be optimally chosen a priori, but rather the
data must be taken sequentially and adaptively in a closed loop. However, unlike
 classification with 0/1 error, collecting data adaptively to find a set with hi
gh true positive rate and low false discovery rate (FDR) is not as well understo
od. In this paper, we provide the first provably sample efficient adaptive algor
ithm for this problem. Along the way, we highlight connections between classific
ation, combinatorial bandits, and FDR control making contributions to each.
****
Online-Within-Online Meta-Learning
Giulia Denevi, Dimitris Stamos, Carlo Ciliberto, Massimiliano Pontil
We study the problem of learning a series of tasks in a fully online Meta-Learni
ng
setting. The goal is to exploit similarities among the tasks to incrementally ad
apt
an inner online algorithm in order to incur a low averaged cumulative error over
the tasks. We focus on a family of inner algorithms based on a parametrized
variant of online Mirror Descent. The inner algorithm is incrementally adapted
by an online Mirror Descent meta-algorithm using the corresponding within-task
minimum regularized empirical risk as the meta-loss. In order to keep the proces
s
fully online, we approximate the meta-subgradients by the online inner algorithm
.
An upper bound on the approximation error allows us to derive a cumulative
error bound for the proposed method. Our analysis can also be converted to the
statistical setting by online-to-batch arguments. We instantiate two examples of
 the
framework in which the meta-parameter is either a common bias vector or feature
map. Finally, preliminary numerical experiments confirm our theoretical findings
.
****
Which Algorithmic Choices Matter at Which Batch Sizes? Insights From a Noisy Qu
adratic Model
Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Da
hl, Chris Shallue, Roger B. Grosse
Increasing the batch size is a popular way to speed up neural network training,
but beyond some critical batch size, larger batch sizes yield diminishing return
s. In this work, we study how the critical batch size changes based on propertie
s of the optimization algorithm, including acceleration and preconditioning, thr
ough two different lenses: large scale experiments and analysis using a simple n
oisy quadratic model (NQM). We experimentally demonstrate that optimization algo
rithms that employ preconditioning, specifically Adam and K-FAC, result in much
larger critical batch sizes than stochastic gradient descent with momentum. We a
lso demonstrate that the NQM captures many of the essential features of real neu
ral network training, despite being drastically simpler to work with. The NQM pr

edicts our results with preconditioned optimizers, previous results with accelerated gradient descent, and other results around optimal learning rates and large batch training, making it a useful tool to generate testable predictions about neural network optimization.
We demonstrate empirically that the simple noisy quadratic model (NQM) displays many similarities to neural networks in terms of large-batch training. We prove analytical convergence results for the NQM model that predict such behavior and hence provide possible explanations and a better understanding for many large-batch training phenomena.
****

Using Statistics to Automate Stochastic Optimization
Hunter Lang, Lin Xiao, Pengchuan Zhang
Despite the development of numerous adaptive optimizers, tuning the learning rate of stochastic gradient methods remains a major roadblock to obtaining good practical performance in machine learning. Rather than changing the learning rate at each iteration, we propose an approach that automates the most common hand-tuning heuristic: use a constant learning rate until "progress stops," then drop. We design an explicit statistical test that determines when the dynamics of stochastic gradient descent reach a stationary distribution. This test can be performed easily during training, and when it fires, we decrease the learning rate by a constant multiplicative factor. Our experiments on several deep learning tasks demonstrate that this statistical adaptive stochastic approximation (SASA) method can automatically find good learning rate schedules and match the performance of hand-tuned methods using default settings of its parameters. The statistical testing helps to control the variance of this procedure and improves its robustness.
****

Margin-Based Generalization Lower Bounds for Boosted Classifiers
Allan Grønlund, Lior Kamma, Kasper Green Larsen, Alexander Mathiasen, Jelani Nelson
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.
****

D-VAE: A Variational Autoencoder for Directed Acyclic Graphs
Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, Yixin Chen
Graph structured data are abundant in the real world. Among different graph types, directed acyclic graphs (DAGs) are of particular interest to machine learning researchers, as many machine learning models are realized as computations on DAGs, including neural networks and Bayesian networks. In this paper, we study deep generative models for DAGs, and propose a novel DAG variational autoencoder (D-VAE). To encode DAGs into the latent space, we leverage graph neural networks. We propose an asynchronous message passing scheme that allows encoding the computations on DAGs, rather than using existing simultaneous message passing schemes to encode local graph structures. We demonstrate the effectiveness of our proposed DVAE through two tasks: neural architecture search and Bayesian network structure learning. Experiments show that our model not only generates novel and valid DAGs, but also produces a smooth latent space that facilitates searching for DAGs with better performance through Bayesian optimization.
****

Adversarial Examples Are Not Bugs, They Are Features
Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry
Adversarial examples have attracted significant attention in machine learning, but the reasons for their existence and pervasiveness remain unclear. We demonstrate that adversarial examples can be directly attributed to the presence of non-robust features: features (derived from patterns in the data distribution) that are highly predictive, yet brittle and (thus) incomprehensible to humans. After capturing these features within a theoretical framework, we establish their wide

spread existence in standard datasets. Finally, we present a simple setting wher e we can rigorously tie the phenomena we observe in practice to a {\em misalignm ent} between the (human-specified) notion of robustness and the inherent geometr y of the data.
****
Characterizing the Exact Behaviors of Temporal Difference Learning Algorithms Us ing Markov Jump Linear System Theory
Bin Hu, Usman Syed
In this paper, we provide a unified analysis of temporal difference learning alg orithms with linear function approximators by exploiting their connections to Ma rkov jump linear systems (MJLS). We tailor the MJLS theory developed in the cont rol community to characterize the exact behaviors of the first and second order moments of a large family of temporal difference learning algorithms. For both t he IID and Markov noise cases, we show that the evolution of some augmented vers ions of the mean and covariance matrix of the TD estimation error exactly follow s the trajectory of a deterministic linear time-invariant (LTI) dynamical system . Applying the well-known LTI system theory, we obtain closed-form expressions f or the mean and covariance matrix of the TD estimation error at any time step. W e provide a tight matrix spectral radius condition to guarantee the convergence of the covariance matrix of the TD estimation error, and perform a perturbation analysis to characterize the dependence of the TD behaviors on learning rate. Fo r the IID case, we provide an exact formula characterizing how the mean and cova riance matrix of the TD estimation error converge to the steady state values at a linear rate. For the Markov case, we use our formulas to explain how the behav iors of TD learning algorithms are affected by learning rate and the underlying Markov chain. For both cases, upper and lower bounds for the mean square TD erro r are provided. The mean square TD error is shown to converge linearly to an exa ct limit.
****
Deep RGB-D Canonical Correlation Analysis For Sparse Depth Completion
Yiqi Zhong, Cho-Ying Wu, Suya You, Ulrich Neumann
In this paper, we propose our Correlation For Completion Network (CFCNet), an en d-to-end deep learning model that uses the correlation between two data sources to perform sparse depth completion. CFCNet learns to capture, to the largest ext ent, the semantically correlated features between RGB and depth information. Thr ough pairs of image pixels and the visible measurements in a sparse depth map, C FCNet facilitates feature-level mutual transformation of different data sources. Such a transformation enables CFCNet to predict features and reconstruct data o f missing depth measurements according to their corresponding, transformed RGB f eatures. We extend canonical correlation analysis to a 2D domain and formulate i t as one of our training objectives (i.e. 2d deep canonical correlation, or “2D^ 2CCA loss"). Extensive experiments validate the ability and flexibility of our CFCNet compared to the state-of-the-art methods on both indoor and outdoor scene s with different real-life sparse patterns. Codes are available at: https://gith ub.com/choyingw/CFCNet.
****
Generalization in Reinforcement Learning with Selective Noise Injection and Info rmation Bottleneck
Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, Katja Hofmann
The ability for policies to generalize to new environments is key to the broad a pplication of RL agents. A promising approach to prevent an agent’s policy from overfitting to a limited set of training environments is to apply regularization techniques originally developed for supervised learning. However, there are sta rk differences between supervised learning and RL. We discuss those differences and propose modifications to existing regularization techniques in order to bett er adapt them to RL. In particular, we focus on regularization techniques relyin g on the injection of noise into the learned function, a family that includes so me of the most widely used approaches such as Dropout and Batch Normalization. T o adapt them to RL, we propose Selective Noise Injection (SNI), which maintains

the regularizing effect the injected noise has, while mitigating the adverse effects it has on the gradient quality. Furthermore, we demonstrate that the Information Bottleneck (IB) is a particularly well suited regularization technique for RL as it is effective in the low-data regime encountered early on in training RL agents. Combining the IB with SNI, we significantly outperform current state of the art results, including on the recently proposed generalization benchmark Coinrun.
****
Untangling in Invariant Speech Recognition
Cory Stephenson, Jenelle Feather, Suchismita Padhy, Oguz Elibol, Hanlin Tang, Josh McDermott, SueYeon Chung
Encouraged by the success of deep convolutional neural networks on a variety of visual tasks, much theoretical and experimental work has been aimed at understanding and interpreting how vision networks operate. At the same time, deep neural networks have also achieved impressive performance in audio processing applications, both as sub-components of larger systems and as complete end-to-end systems by themselves. Despite their empirical successes, comparatively little is understood about how these audio models accomplish these tasks.In this work, we employ a recently developed statistical mechanical theory that connects geometric properties of network representations and the separability of classes to probe how information is untangled within neural networks trained to recognize speech. We observe that speaker-specific nuisance variations are discarded by the network's hierarchy, whereas task-relevant properties such as words and phonemes are untangled in later layers. Higher level concepts such as parts-of-speech and context dependence also emerge in the later layers of the network. Finally, we find that the deep representations carry out significant temporal untangling by efficiently extracting task-relevant features at each time step of the computation. Taken together, these findings shed light on how deep auditory models process their time dependent input signals to carry out invariant speech recognition, and show how different concepts emerge through the layers of the network.
****
Fast structure learning with modular regularization
Greg Ver Steeg, Hrayr Harutyunyan, Daniel Moyer, Aram Galstyan
Estimating graphical model structure from high-dimensional and undersampled data is a fundamental problem in many scientific fields.
Existing approaches, such as GLASSO, latent variable GLASSO, and latent tree models, suffer from high computational complexity and may impose unrealistic sparsity priors in some cases.
We introduce a novel method that leverages a newly discovered connection between information-theoretic measures and structured latent factor models to derive an optimization objective which encourages modular structures where each observed variable has a single latent parent.
The proposed method has linear stepwise computational complexity w.r.t. the number of observed variables.
Our experiments on synthetic data demonstrate that our approach is the only method that recovers modular structure better as the dimensionality increases. We also use our approach for estimating covariance structure for a number of real-world datasets and show that it consistently outperforms state-of-the-art estimators at a fraction of the computational cost. Finally, we apply the proposed method to high-resolution fMRI data (with more than 10^5 voxels) and show that it is capable of extracting meaningful patterns.
****
Graph-based Discriminators: Sample Complexity and Expressiveness
Roi Livni, Yishay Mansour
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.
****
Certifiable Robustness to Graph Perturbations

Aleksandar Bojchevski, Stephan Günnemann
Despite the exploding interest in graph neural networks there has been little ef fort to verify and improve their robustness. This is even more alarming given re cent findings showing that they are extremely vulnerable to adversarial attacks on both the graph structure and the node attributes. We propose the first method for verifying certifiable (non-)robustness to graph perturbations for a general class of models that includes graph neural networks and label/feature propagati on. By exploiting connections to PageRank and Markov decision processes our cert ificates can be efficiently (and under many threat models exactly) computed. Fur thermore, we investigate robust training procedures that increase the number of certifiably robust nodes while maintaining or improving the clean predictive acc uracy.
****
Surfing: Iterative Optimization Over Incrementally Trained Deep Networks
Ganlin Song, Zhou Fan, John Lafferty
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.
****
Rates of Convergence for Large-scale Nearest Neighbor Classification
Xingye Qiao, Jiexin Duan, Guang Cheng
Nearest neighbor is a popular class of classification methods with many desirabl e properties. For a large data set which cannot be loaded into the memory of a s ingle machine due to computation, communication, privacy, or ownership limitatio ns, we consider the divide and conquer scheme: the entire data set is divided in to small subsamples, on which nearest neighbor predictions are made, and then a final decision is reached by aggregating the predictions on subsamples by majori ty voting. We name this method the big Nearest Neighbor (bigNN) classifier, and provide its rates of convergence under minimal assumptions, in terms of both the excess risk and the classification instability, which are proven to be the same rates as the oracle nearest neighbor classifier and cannot be improved. To sign ificantly reduce the prediction time that is required for achieving the optimal rate, we also consider the pre-training acceleration technique applied to the bi gNN method, with proven convergence rate. We find that in the distributed settin g, the optimal choice of the neighbor k should scale with both the total sample size and the number of partitions, and there is a theoretical upper limit for th e latter. Numerical studies have verified the theoretical findings.
****
Finite-Time Performance Bounds and Adaptive Learning Rate Selection for Two Time -Scale Reinforcement Learning
Harsh Gupta, R. Srikant, Lei Ying
We study two time-scale linear stochastic approximation algorithms, which can be used to model well-known reinforcement learning algorithms such as GTD, GTD2, a nd TDC. We present finite-time performance bounds for the case where the learnin g rate is fixed. The key idea in obtaining these bounds is to use a Lyapunov fun ction motivated by singular perturbation theory for linear differential equation s. We use the bound to design an adaptive learning rate scheme which significant ly improves the convergence rate over the known optimal polynomial decay rule in our experiments, and can be used to potentially improve the performance of any other schedule where the learning rate is changed at pre-determined time instant s.
****
Pseudo-Extended Markov chain Monte Carlo
Christopher Nemeth, Fredrik Lindsten, Maurizio Filippone, James Hensman
Sampling from posterior distributions using Markov chain Monte Carlo (MCMC) meth ods can require an exhaustive number of iterations, particularly when the poster ior is multi-modal as the MCMC sampler can become trapped in a local mode for a large number of iterations. In this paper, we introduce the pseudo-extended MCMC method as a simple approach for improving the mixing of the MCMC sampler for mu

lti-modal posterior distributions. The pseudo-extended method augments the state-space of the posterior using pseudo-samples as auxiliary variables. On the extended space, the modes of the posterior are connected, which allows the MCMC sampler to easily move between well-separated posterior modes. We demonstrate that the pseudo-extended approach delivers improved MCMC sampling over the Hamiltonian Monte Carlo algorithm on multi-modal posteriors, including Boltzmann machines and models with sparsity-inducing priors.
****
Hierarchical Optimal Transport for Multimodal Distribution Alignment
John Lee, Max Dabagia, Eva Dyer, Christopher Rozell
In many machine learning applications, it is necessary to meaningfully aggregate, through alignment, different but related datasets. Optimal transport (OT)-based approaches pose alignment as a divergence minimization problem: the aim is to transform a source dataset to match a target dataset using the Wasserstein distance as a divergence measure. We introduce a hierarchical formulation of OT which leverages clustered structure in data to improve alignment in noisy, ambiguous, or multimodal settings. To solve this numerically, we propose a distributed ADMM algorithm that also exploits the Sinkhorn distance, thus it has an efficient computational complexity that scales quadratically with the size of the largest cluster. When the transformation between two datasets is unitary, we provide performance guarantees that describe when and how well aligned cluster correspondences can be recovered with our formulation, as well as provide worst-case dataset geometry for such a strategy. We apply this method to synthetic datasets that model data as mixtures of low-rank Gaussians and study the impact that different geometric properties of the data have on alignment. Next, we applied our approach to a neural decoding application where the goal is to predict movement directions and instantaneous velocities from populations of neurons in the macaque primary motor cortex. Our results demonstrate that when clustered structure exists in datasets, and is consistent across trials or time points, a hierarchical alignment strategy that leverages such structure can provide significant improvements in cross-domain alignment.
****
Sampled Softmax with Random Fourier Features
Ankit Singh Rawat, Jiecao Chen, Felix Xinnan X. Yu, Ananda Theertha Suresh, Sanjiv Kumar
The computational cost of training with softmax cross entropy loss grows linearly with the number of classes. For the settings where a large number of classes are involved, a common method to speed up training is to sample a subset of classes and utilize an estimate of the loss gradient based on these classes, known as the sampled softmax method. However, the sampled softmax provides a biased estimate of the gradient unless the samples are drawn from the exact softmax distribution, which is again expensive to compute. Therefore, a widely employed practical approach involves sampling from a simpler distribution in the hope of approximating the exact softmax distribution. In this paper, we develop the first theoretical understanding of the role that different sampling distributions play in determining the quality of sampled softmax. Motivated by our analysis and the work on kernel-based sampling, we propose the Random Fourier Softmax (RF-softmax) method that utilizes the powerful Random Fourier Features to enable more efficient and accurate sampling from an approximate softmax distribution. We show that RF-softmax leads to low bias in estimation in terms of both the full softmax distribution and the full softmax gradient. Furthermore, the cost of RF-softmax scales only logarithmically with the number of classes.
****
Epsilon-Best-Arm Identification in Pay-Per-Reward Multi-Armed Bandits
Sivan Sabato
We study epsilon-best-arm identification, in a setting where during the exploration phase, the cost of each arm pull is proportional to the expected future reward of that arm. We term this setting Pay-Per-Reward. We provide an algorithm for this setting, that with a high probability returns an epsilon-best arm, while incurring a cost that depends only linearly on the total expected reward of all

arms, and does not depend at all on the number of arms. Under mild assumptions, the algorithm can be applied also to problems with infinitely many arms.
****
On Differentially Private Graph Sparsification and Applications
Raman Arora, Jalaj Upadhyay
In this paper, we study private sparsification of graphs. In particular, we give an algorithm that given an input graph, returns a sparse graph which approximates the spectrum of the input graph while ensuring differential privacy. This allows one to solve many graph problems privately yet efficiently and accurately. This is exemplified with application of the proposed meta-algorithm to graph algorithms for privately answering cut-queries, as well as practical algorithms for computing {\scshape MAX-CUT} and {\scshape SPARSEST-CUT} with better accuracy than previously known. We also give the first efficient private algorithm to learn Laplacian eigenmap on a graph.
****
On the number of variables to use in principal component regression
Ji Xu, Daniel J. Hsu
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.
****
Self-Routing Capsule Networks
Taeyoung Hahn, Myeongjang Pyeon, Gunhee Kim
Capsule networks have recently gained a great deal of interest as a new architecture of neural networks that can be more robust to input perturbations than similar-sized CNNs. Capsule networks have two major distinctions from the conventional CNNs: (i) each layer consists of a set of capsules that specialize in disjoint regions of the feature space and (ii) the routing-by-agreement coordinates connections between adjacent capsule layers. Although the routing-by-agreement is capable of filtering out noisy predictions of capsules by dynamically adjusting their influences, its unsupervised clustering nature causes two weaknesses: (i) high computational complexity and (ii) cluster assumption that may not hold in presence of heavy input noise. In this work, we propose a novel and surprisingly simple routing strategy called self-routing where each capsule is routed independently by its subordinate routing network. Therefore, the agreement between capsules is not required anymore but both poses and activations of upper-level capsules are obtained in a way similar to Mixture-of-Experts. Our experiments on CIFAR-10, SVHN and SmallNORB show that the self-routing performs more robustly against white-box adversarial attacks and affine transformations, requiring less computation.
****
A Model-Based Reinforcement Learning with Adversarial Training for Online Recommendation
Xueying Bai, Jian Guan, Hongning Wang
Reinforcement learning is effective in optimizing policies for recommender systems. Current solutions mostly focus on model-free approaches, which require frequent interactions with a real environment, and thus are expensive in model learning. Offline evaluation methods, such as importance sampling, can alleviate such limitations, but usually request a large amount of logged data and do not work well when the action space is large. In this work, we propose a model-based reinforcement learning solution which models the user-agent interaction for offline policy learning via a generative adversarial network. To reduce bias in the learnt policy, we use the discriminator to evaluate the quality of generated sequences and rescale the generated rewards. Our theoretical analysis and empirical evaluations demonstrate the effectiveness of our solution in identifying patterns from given offline data and learning policies based on the offline and generated data.
****
Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation

Risto Vuorio, Shao-Hua Sun, Hexiang Hu, Joseph J. Lim
Model-agnostic meta-learners aim to acquire meta-learned parameters from similar tasks to adapt to novel tasks from the same distribution with few gradient updates. With the flexibility in the choice of models, those frameworks demonstrate appealing performance on a variety of domains such as few-shot image classification and reinforcement learning. However, one important limitation of such frameworks is that they seek a common initialization shared across the entire task distribution, substantially limiting the diversity of the task distributions that they are able to learn from. In this paper, we augment MAML with the capability to identify the mode of tasks sampled from a multimodal task distribution and adapt quickly through gradient updates. Specifically, we propose a multimodal MAML (MMAML) framework, which is able to modulate its meta-learned prior parameters according to the identified mode, allowing more efficient fast adaptation. We evaluate the proposed model on a diverse set of few-shot learning tasks, including regression, image classification, and reinforcement learning. The results not only demonstrate the effectiveness of our model in modulating the meta-learned prior in response to the characteristics of tasks but also show that training on a multimodal distribution can produce an improvement over unimodal training. The code for this project is publicly available at https://vuoristo.github.io/MMAML.
****
Predicting the Politics of an Image Using Webly Supervised Data
Christopher Thomas, Adriana Kovashka
The news media shape public opinion, and often, the visual bias they contain is evident for human observers. This bias can be inferred from how different media sources portray different subjects or topics. In this paper, we model visual political bias in contemporary media sources at scale, using webly supervised data. We collect a dataset of over one million unique images and associated news articles from left- and right-leaning news sources, and develop a method to predict the image's political leaning. This problem is particularly challenging because of the enormous intra-class visual and semantic diversity of our data. We propose a two-stage method to tackle this problem. In the first stage, the model is forced to learn relevant visual concepts that, when joined with document embeddings computed from articles paired with the images, enable the model to predict bias. In the second stage, we remove the requirement of the text domain and train a visual classifier from the features of the former model. We show this two-stage approach facilitates learning and outperforms several strong baselines. We also present extensive qualitative results demonstrating the nuances of the data.
****
On the Curved Geometry of Accelerated Optimization
Aaron Defazio
In this work we propose a differential geometric motivation for Nesterov's accelerated gradient method (AGM) for strongly-convex problems. By considering the optimization procedure as occurring on a Riemannian manifold with a natural structure, The AGM method can be seen as the proximal point method applied in this curved space. This viewpoint can also be extended to the continuous time case, where the accelerated gradient method arises from the natural block-implicit Euler discretization of an ODE on the manifold. We provide an analysis of the convergence rate of this ODE for quadratic objectives.
****
How to Initialize your Network? Robust Initialization for WeightNorm & ResNets
Devansh Arpit, Víctor Campos, Yoshua Bengio
Residual networks (ResNet) and weight normalization play an important role in various deep learning applications. However, parameter initialization strategies have not been studied previously for weight normalized networks and, in practice, initialization methods designed for un-normalized networks are used as a proxy. Similarly, initialization for ResNets have also been studied for un-normalized networks and often under simplified settings ignoring the shortcut connection. To address these issues, we propose a novel parameter initialization strategy that avoids explosion/vanishment of information across layers for weight normalized networks with and without residual connections. The proposed strategy is based

on a theoretical analysis using mean field approximation. We run over 2,500 expe riments and evaluate our proposal on image datasets showing that the proposed in itialization outperforms existing initialization methods in terms of generalizat ion performance, robustness to hyper-parameter values and variance between seeds , especially when networks get deeper in which case existing methods fail to eve n start training. Finally, we show that using our initialization in conjunction with learning rate warmup is able to reduce the gap between the performance of w eight normalized and batch normalized networks.
****
Code Generation as a Dual Task of Code Summarization
Bolin Wei, Ge Li, Xin Xia, Zhiyi Fu, Zhi Jin
Code summarization (CS) and code generation (CG) are two crucial tasks in the fi eld of automatic software development. Various neural network-based approaches a re proposed to solve these two tasks separately. However, there exists a specifi c intuitive correlation between CS and CG, which has not been exploited in previ ous work. In this paper, we apply the relations between two tasks to improve the performance of both tasks. In other words, exploiting the duality between the t wo tasks, we propose a dual training framework to train the two tasks simultaneo usly. In this framework, we consider the dualities on probability and attention weights, and design corresponding regularization terms to constrain the duality. We evaluate our approach on two datasets collected from GitHub, and experimenta l results show that our dual framework can improve the performance of CS and CG tasks over baselines.
****
A Graph Theoretic Framework of Recomputation Algorithms for Memory-Efficient Bac kpropagation
Mitsuru Kusumoto, Takuya Inoue, Gentaro Watanabe, Takuya Akiba, Masanori Koyama
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.
****
Gradient based sample selection for online continual learning
Rahaf Aljundi, Min Lin, Baptiste Goujaud, Yoshua Bengio
A continual learning agent learns online with a non-stationary and never-ending stream of data. The key to such learning process is to overcome the catastrophic forgetting of previously seen data, which is a well known problem of neural net works. To prevent forgetting, a replay buffer is usually employed to store the p revious data for the purpose of rehearsal. Previous work often depend on task bo undary and i.i.d. assumptions to properly select samples for the replay buffer. In this work, we formulate sample selection as a constraint reduction problem ba sed on the constrained optimization view of continual learning. The goal is to s elect a fixed subset of constraints that best approximate the feasible region de fined by the original constraints. We show that it is equivalent to maximizing t he diversity of samples in the replay buffer with parameter gradient as the feat ure. We further develop a greedy alternative that is cheap and efficient. The ad vantage of the proposed method is demonstrated by comparing to other alternative s under the continual learning setting. Further comparisons are made against sta te of the art methods that rely on task boundaries which show comparable or even better results for our method.
****
Conditional Structure Generation through Graph Variational Generative Adversaria l Nets
Carl Yang, Peiye Zhuang, Wenhan Shi, Alan Luu, Pan Li
Graph embedding has been intensively studied recently, due to the advance of var ious neural network models. Theoretical analyses and empirical studies have push ed forward the translation of discrete graph structures into distributed represe ntation vectors, but seldom considered the reverse direction, i.e., generation o f graphs from given related context spaces. Particularly, since graphs often bec ome more meaningful when associated with semantic contexts (e.g., social network

s of certain communities, gene networks of certain diseases), the ability to inf
er graph structures according to given semantic conditions could be of great val
ue. While existing graph generative models only consider graph structures withou
t semantic contexts, we formulate the novel problem of conditional structure gen
eration, and propose a novel unified model of graph variational generative adver
sarial nets (CondGen) to handle the intrinsic challenges of flexible context-str
ucture conditioning and permutation-invariant generation. Extensive experiments
on two deliberately created benchmark datasets of real-world context-enriched ne
tworks demonstrate the supreme effectiveness and generalizability of CondGen.
****
Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting
Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, Deyu Meng
Current deep neural networks(DNNs) can easily overfit to biased training data wi
th corrupted labels or class imbalance. Sample re-weighting strategy is commonly
 used to alleviate this issue by designing a weighting function mapping from tra
ining loss to sample weight, and then iterating between weight recalculating and
 classifier updating. Current approaches, however, need manually pre-specify the
 weighting function as well as its additional hyper-parameters. It makes them fa
irly hard to be generally applied in practice due to the significant variation o
f proper weighting schemes relying on the investigated problem and training data
. To address this issue, we propose a method capable of adaptively learning an e
xplicit weighting function directly from data. The weighting function is an MLP
with one hidden layer, constituting a universal approximator to almost any conti
nuous functions, making the method able to fit a wide range of weighting functio
n forms including those assumed in conventional research. Guided by a small amou
nt of unbiased meta-data, the parameters of the weighting function can be finely
 updated simultaneously with the learning process of the classifiers. Synthetic
and real experiments substantiate the capability of our method for achieving pro
per weighting functions in class imbalance and noisy label cases, fully complyin
g with the common settings in traditional methods, and more complicated scenario
s beyond conventional cases. This naturally leads to its better accuracy than ot
her state-of-the-art methods.
****
Thompson Sampling with Information Relaxation Penalties
Seungki Min, Costis Maglaras, Ciamac C. Moallemi
We consider a finite-horizon multi-armed bandit (MAB) problem in a Bayesian sett
ing, for which we propose an information relaxation sampling framework. With thi
s framework, we define an intuitive family of control policies that include Thom
pson sampling (TS) and the Bayesian optimal policy as endpoints. Analogous to TS
, which, at each decision epoch pulls an arm that is best with respect to the ra
ndomly sampled parameters, our algorithms sample entire future reward realizatio
ns and take the corresponding best action. However, this is done in the presence
 of “penalties” that seek to compensate for the availability of future informati
on.
****
Oracle-Efficient Algorithms for Online Linear Optimization with Bandit Feedback
Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori
 Kakimura, Ken-Ichi Kawarabayashi
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.
****
Constraint-based Causal Structure Learning with Consistent Separating Sets
Honghao Li, Vincent Cabeli, Nadir Sella, Herve Isambert
We consider constraint-based methods for causal structure learning, such as the
PC algorithm or any PC-derived algorithms whose ■rst step consists in pruning a
complete graph to obtain an undirected graph skeleton, which is subsequently ori
ented. All constraint-based methods perform this ■rst step of removing dispensab
le edges, iteratively, whenever a separating set and corresponding conditional i

ndependence can be found. Yet, constraint-based methods lack robustness over sam
pling noise and are prone to uncover spurious conditional independences in ■nite
 datasets. In particular, there is no guarantee that the separating sets identi■
ed during the iterative pruning step remain consistent with the ■nal graph. In t
his paper, we propose a simple modi■cation of PC and PC-derived algorithms so as
 to ensure that all separating sets identi■ed to remove dispensable edges are co
nsistent with the ■nal graph,thus enhancing the explainability of constraint-bas
edmethods. It is achieved by repeating the constraint-based causal structure lea
rning scheme, iteratively, while searching for separating sets that are consiste
nt with the graph obtained at the previous iteration. Ensuring the consistency o
f separating sets can be done at a limited complexity cost, through the use of b
lock-cut tree decomposition of graph skeletons, and is found to increase their v
alidity in terms of actual d-separation. It also signi■cantly improves the sensi
tivity of constraint-based methods while retaining good overall structure learni
ng performance. Finally and foremost, ensuring sepset consistency improves the i
nterpretability of constraint-based models for real-life applications.
****
Efficient Deep Approximation of GMMs
Shirin Jalali, Carl Nuzman, Iraj Saniee
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.
****
Selecting causal brain features with a single conditional independence test per
feature
Atalanti Mastakouri, Bernhard Schölkopf, Dominik Janzing
We propose a constraint-based causal feature selection method for identifying ca
uses of a given target variable, selecting from a set of candidate variables, wh
ile there can also be hidden variables acting as common causes with the target.
We prove that if we observe a cause for each candidate cause, then a single cond
itional independence test with one conditioning variable is sufficient to decide
 whether a candidate associated with the target is indeed causing it. We thus im
prove upon existing methods by significantly simplifying statistical testing and
 requiring a weaker version of causal faithfulness. Our main assumption is inspi
red by neuroscience paradigms where the activity of a single neuron is considere
d to be also caused by its own previous state. We demonstrate successful applica
tion of our method to simulated, as well as encephalographic data of twenty-one
participants, recorded in Max Planck Institute for intelligent Systems. The det
ected causes of motor performance are in accordance with the latest consensus ab
out the neurophysiological pathways, and can provide new insights into personali
sed brain stimulation.
****
Sample-Efficient Deep Reinforcement Learning via Episodic Backward Update
Su Young Lee, Choi Sungik, Sae-Young Chung
We propose Episodic Backward Update (EBU) – a novel deep reinforcement learning
algorithm with a direct value propagation. In contrast to the conventional use o
f the experience replay with uniform random sampling, our agent samples a whole
episode and successively propagates the value of a state to its previous states.
 Our computationally efficient recursive algorithm allows sparse and delayed rew
ards to propagate directly through all transitions of the sampled episode. We th
eoretically prove the convergence of the EBU method and experimentally demonstra
te its performance in both deterministic and stochastic environments. Especially
 in 49 games of Atari 2600 domain, EBU achieves the same mean and median human n
ormalized performance of DQN by using only 5% and 10% of samples, respectively.
****
Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior
Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, Yung-Yu Chuang
This paper presents a weakly supervised instance segmentation method that consum
es training data with tight bounding box annotations. The major difficulty lies

in the uncertain figure-ground separation within each bounding box since there is no supervisory signal about it. We address the difficulty by formulating the problem as a multiple instance learning (MIL) task, and generate positive and negative bags based on the sweeping lines of each bounding box. The proposed deep model integrates MIL into a fully supervised instance segmentation network, and can be derived by the objective consisting of two terms, i.e., the unary term and the pairwise term. The former estimates the foreground and background areas of each bounding box while the latter maintains the unity of the estimated object masks. The experimental results show that our method performs favorably against existing weakly supervised methods and even surpasses some fully supervised methods for instance segmentation on the PASCAL VOC dataset.
****

Copula-like Variational Inference

Marcel Hirt, Petros Dellaportas, Alain Durmus

This paper considers a new family of variational distributions motivated by Sklar's theorem. This family is based on new copula-like densities on the hypercube with non-uniform marginals which can be sampled efficiently, i.e. with a complexity linear in the dimension d of the state space. Then, the proposed variational densities that we suggest can be seen as arising from these copula-like densities used as base distributions on the hypercube with Gaussian quantile functions and sparse rotation matrices as normalizing flows. The latter correspond to a rotation of the marginals with complexity O(d log d). We provide some empirical evidence that such a variational family can also approximate non-Gaussian posteriors and can be beneficial compared to Gaussian approximations. Our method performs largely comparably to state-of-the-art variational approximations on standard regression and classification benchmarks for Bayesian Neural Networks.
****

Towards Hardware-Aware Tractable Learning of Probabilistic Models

Laura I. Galindez Olascoaga, Wannes Meert, Nimish Shah, Marian Verhelst, Guy Van den Broeck

Smart portable applications increasingly rely on edge computing due to privacy and latency concerns. But guaranteeing always-on functionality comes with two major challenges: heavily resource-constrained hardware; and dynamic application conditions. Probabilistic models present an ideal solution to these challenges: they are robust to missing data, allow for joint predictions and have small data needs. In addition, ongoing efforts in field of tractable learning have resulted in probabilistic models with strict inference efficiency guarantees. However, the current notions of tractability are often limited to model complexity, disregarding the hardware's specifications and constraints. We propose a novel resource-aware cost metric that takes into consideration the hardware's properties in determining whether the inference task can be efficiently deployed. We use this metric to evaluate the performance versus resource trade-off relevant to the application of interest, and we propose a strategy that selects the device-settings that can optimally meet users' requirements. We showcase our framework on a mobile activity recognition scenario, and on a variety of benchmark datasets representative of the field of tractable learning and of the applications of interest.
****

Two Time-scale Off-Policy TD Learning: Non-asymptotic Analysis over Markovian Samples

Tengyu Xu, Shaofeng Zou, Yingbin Liang

Gradient-based temporal difference (GTD) algorithms are widely used in off-policy learning scenarios. Among them, the two time-scale TD with gradient correction (TDC) algorithm has been shown to have superior performance. In contrast to previous studies that characterized the non-asymptotic convergence rate of TDC only under identical and independently distributed (i.i.d.) data samples, we provide the first non-asymptotic convergence analysis for two time-scale TDC under a non-i.i.d.\ Markovian sample path and linear function approximation. We show that the two time-scale TDC can converge as fast as O(log t/t^(2/3)) under diminishing stepsize, and can converge exponentially fast under constant stepsize, but at the cost of a non-vanishing error. We further propose a TDC algorithm with block

wisely diminishing stepsize, and show that it asymptotically converges with an a
rbitrarily small error at a blockwisely linear convergence rate. Our experiments
 demonstrate that such an algorithm converges as fast as TDC under constant step
size, and still enjoys comparable accuracy as TDC under diminishing stepsize.
****
Incremental Few-Shot Learning with Attention Attractor Networks
Mengye Ren, Renjie Liao, Ethan Fetaya, Richard Zemel
Machine learning classifiers are often trained to recognize a set of pre-defined
 classes. However, in many applications, it is often desirable to have the flexi
bility of learning additional concepts, with limited data and without re-trainin
g on the full training set. This paper addresses this problem, incremental few-s
hot learning, where a regular classification network has already been trained to
 recognize a set of base classes, and several extra novel classes are being cons
idered, each with only a few labeled examples. After learning the novel classes,
 the model is then evaluated on the overall classification performance on both b
ase and novel classes. To this end, we propose a meta-learning model, the Attent
ion Attractor Network, which regularizes the learning of novel classes. In each
episode, we train a set of new weights to recognize novel classes until they con
verge, and we show that the technique of recurrent back-propagation can back-pro
pagate through the optimization process and facilitate the learning of these par
ameters. We demonstrate that the learned attractor network can help recognize no
vel classes while remembering old classes without the need to review the origina
l training set, outperforming various baselines.
****
Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Ob
ject Representations
Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbau
m, Tomer Ullman
From infancy, humans have expectations about how objects will move and interact.
 Even young children expect objects not to move through one another, teleport, o
r disappear. They are surprised by mismatches between physical expectations and
perceptual observations, even in unfamiliar scenes with completely novel objects
. A model that exhibits human-like understanding of physics should be similarly
surprised, and adjust its beliefs accordingly. We propose ADEPT, a model that u
ses a coarse (approximate geometry) object-centric representation for dynamic 3D
 scene understanding. Inference integrates deep recognition networks, extended p
robabilistic physical simulation, and particle filtering for forming predictions
 and expectations across occlusion. We also present a new test set for measuring
 violations of physical expectations, using a range of scenarios derived from de
velopmental psychology. We systematically compare ADEPT, baseline models, and h
uman expectations on this test set. ADEPT outperforms standard network architec
tures in discriminating physically implausible scenes, and often performs this d
iscrimination at the same level as people.
****
Outlier Detection and Robust PCA Using a Convex Measure of Innovation
Mostafa Rahmani, Ping Li
This paper presents a provable and strong algorithm, termed Innovation Search (
iSearch), to robust Principal Component Analysis (PCA) and outlier detection. A
n outlier by definition is a data point which does not participate in forming a
 low dimensional structure with a large number of data points in the data. In o
ther word, an outlier carries some innovation with respect to most of the other
data points. iSearch ranks the data points based on their values of innovation.
A convex optimization problem is proposed whose optimal value is used as our mea
sure of innovation. We derive analytical performance guarantees for the propos
ed robust PCA method under different models for the distribution of the outlier
s including randomly distributed outliers, clustered outliers, and linearly depe
ndent outliers. Moreover, it is shown that iSearch provably recovers the span o
f the inliers when the inliers lie in a union of subspaces. In the challenging
scenarios in which the outliers are close to each other or they are close to the
 span of the inliers, iSearch is shown to outperform most of the existing metho

ds.
****
Efficient Convex Relaxations for Streaming PCA
Raman Arora, Teodor Vanislavov Marinov
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.
****
Envy-Free Classification
Maria-Florina F. Balcan, Travis Dick, Ritesh Noothigattu, Ariel D. Procaccia
In classic fair division problems such as cake cutting and rent division, envy-freeness requires that each individual (weakly) prefer his allocation to anyone else's. On a conceptual level, we argue that envy-freeness also provides a compelling notion of fairness for classification tasks, especially when individuals have heterogeneous preferences. Our technical focus is the generalizability of envy-free classification, i.e., understanding whether a classifier that is envy free on a sample would be almost envy free with respect to the underlying distribution with high probability. Our main result establishes that a small sample is sufficient to achieve such guarantees, when the classifier in question is a mixture of deterministic classifiers that belong to a family of low Natarajan dimension.
****
Deep Model Transferability from Attribution Maps
Jie Song, Yixin Chen, Xinchao Wang, Chengchao Shen, Mingli Song
Exploring the transferability between heterogeneous tasks sheds light on their intrinsic interconnections, and consequently enables knowledge transfer from one task to another so as to reduce the training effort of the latter. In this paper, we propose an embarrassingly simple yet very efficacious approach to estimating the transferability of deep networks, especially those handling vision tasks. Unlike the seminal work of \emph{taskonomy} that relies on a large number of annotations as supervision and is thus computationally cumbersome, the proposed approach requires no human annotations and imposes no constraints on the architectures of the networks. This is achieved, specifically, via projecting deep networks into a \emph{model space}, wherein each network is treated as a point and the distances between two points are measured by deviations of their produced attribution maps. The proposed approach is several-magnitude times faster than taskonomy, and meanwhile preserves a task-wise topological structure highly similar to the one obtained by taskonomy. Code is available at \url{https://github.com/zju-vipa/TransferbilityFromAttributionMaps}.
****
Towards Interpretable Reinforcement Learning Using Attention Augmented Agents
Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, Danilo Jimenez Rezende
Inspired by recent work in attention models for image captioning and question answering, we present a soft attention model for the reinforcement learning domain. This model bottlenecks the view of an agent by a soft, top-down attention mechanism, forcing the agent to focus on task-relevant information by sequentially querying its view of the environment. The output of the attention mechanism allows direct observation of the information used by the agent to select its actions, enabling easier interpretation of this model than of traditional models. We analyze the different strategies the agents learn and show that a handful of strategies arise repeatedly across different games. We also show that the model learns to query separately about space and content (where'' vs.what'').
We demonstrate that an agent using this mechanism can achieve performance competitive with state-of-the-art models on ATARI tasks while still being interpretable.
****
Weight Agnostic Neural Networks
Adam Gaier, David Ha

Not all neural network architectures are created equal, some perform much better than others for certain tasks. But how important are the weight parameters of a neural network compared to its architecture? In this work, we question to what extent neural network architectures alone, without learning any weight parameters, can encode solutions for a given task. We propose a search method for neural network architectures that can already perform a task without any explicit weight training. To evaluate these networks, we populate the connections with a single shared weight parameter sampled from a uniform random distribution, and measure the expected performance. We demonstrate that our method can find minimal neural network architectures that can perform several reinforcement learning tasks without weight training. On a supervised learning domain, we find network architectures that achieve much higher than chance accuracy on MNIST using random weights.
****
DeepWave: A Recurrent Neural-Network for Real-Time Acoustic Imaging
Matthieu SIMEONI, Sepand Kashani, Paul Hurley, Martin Vetterli
We propose a recurrent neural-network for real-time reconstruction of acoustic camera spherical maps. The network, dubbed DeepWave, is both physically and algorithmically motivated: its recurrent architecture mimics iterative solvers from convex optimisation, and its parsimonious parametrisation is based on the natural structure of acoustic imaging problems.
Each network layer applies successive filtering, biasing and activation steps to its input, which can be interpreted as generalised deblurring and sparsification steps. To comply with the irregular geometry of spherical maps, filtering operations are implemented efficiently by means of graph signal processing techniques.
Unlike commonly-used imaging network architectures, DeepWave is moreover capable of directly processing the complex-valued raw microphone correlations, learning how to optimally back-project these into a spherical map. We propose moreover a smart physically-inspired initialisation scheme that attains much faster training and higher performance than random initialisation.
Our real-data experiments show DeepWave has similar computational speed to the state-of-the-art delay-and-sum imager with vastly superior resolution. While developed primarily for acoustic cameras, DeepWave could easily be adapted to neighbouring signal processing fields, such as radio astronomy, radar and sonar.
****
A Linearly Convergent Proximal Gradient Algorithm for Decentralized Optimization
Sulaiman Alghunaim, Kun Yuan, Ali H. Sayed
Decentralized optimization is a powerful paradigm that finds applications in engineering and learning design. This work studies decentralized composite optimization problems with non-smooth regularization terms. Most existing gradient-based proximal decentralized methods are known to converge to the optimal solution with sublinear rates, and it remains unclear whether this family of methods can achieve global linear convergence. To tackle this problem, this work assumes the non-smooth regularization term is common across all networked agents, which is the case for many machine learning problems. Under this condition, we design a proximal gradient decentralized algorithm whose fixed point coincides with the desired minimizer. We then provide a concise proof that establishes its linear convergence. In the absence of the non-smooth term, our analysis technique covers the well known EXTRA algorithm and provides useful bounds on the convergence rate and step-size.
****
Meta Architecture Search
Albert Shaw, Wei Wei, Weiyang Liu, Le Song, Bo Dai
Neural Architecture Search (NAS) has been quite successful in constructing state-of-the-art models on a variety of tasks. Unfortunately, the computational cost can make it difficult to scale. In this paper, we make the first attempt to study Meta Architecture Search which aims at learning a task-agnostic representation that can be used to speed up the process of architecture search on a large numb

er of tasks. We propose the Bayesian Meta Architecture SEarch (BASE) framework which takes advantage of a Bayesian formulation of the architecture search problem to learn over an entire set of tasks simultaneously. We show that on Imagenet classification, we can find a model that achieves 25.7% top-1 error and 8.1% top-5 error by adapting the architecture in less than an hour from an 8 GPU days pretrained meta-network. By learning a good prior for NAS, our method dramatically decreases the required computation cost while achieving comparable performance to current state-of-the-art methods - even finding competitive models for unseen datasets with very quick adaptation. We believe our framework will open up new possibilities for efficient and massively scalable architecture search research across multiple tasks.
****
Double Quantization for Communication-Efficient Distributed Optimization
Yue Yu, Jiaxiang Wu, Longbo Huang
Modern distributed training of machine learning models often suffers from high communication overhead for synchronizing stochastic gradients and model parameters. In this paper, to reduce the communication complexity, we propose \emph{double quantization}, a general scheme for quantizing both model parameters and gradients. Three communication-efficient algorithms are proposed based on this general scheme. Specifically, (i) we propose a low-precision algorithm AsyLPG with asynchronous parallelism, (ii) we explore integrating gradient sparsification with double quantization and develop Sparse-AsyLPG, (iii) we show that double quantization can be accelerated by the momentum technique and design accelerated AsyLPG. We establish rigorous performance guarantees for the algorithms, and conduct experiments on a multi-server test-bed with real-world datasets to demonstrate that our algorithms can effectively save transmitted bits without performance degradation, and significantly outperform existing methods with either model parameter or gradient quantization.
****
Graph Structured Prediction Energy Networks
Colin Graber, Alexander Schwing
For joint inference over multiple variables, a variety of structured prediction techniques have been developed to model correlations among variables and thereby improve predictions. However, many classical approaches suffer from one of two primary drawbacks: they either lack the ability to model high-order correlations among variables while maintaining computationally tractable inference, or they do not allow to explicitly model known correlations. To address this shortcoming, we introduce ‘Graph Structured Prediction Energy Networks,’ for which we develop inference techniques that allow to both model explicit local and implicit higher-order correlations while maintaining tractability of inference. We apply the proposed method to tasks from the natural language processing and computer vision domain and demonstrate its general utility.
****
Universal Invariant and Equivariant Graph Neural Networks
Nicolas Keriven, Gabriel Peyré
Graph Neural Networks (GNN) come in many flavors, but should always be either invariant (permutation of the nodes of the input graph does not affect the output) or \emph{equivariant} (permutation of the input permutes the output). In this paper, we consider a specific class of invariant and equivariant networks, for which we prove new universality theorems. More precisely, we consider networks with a single hidden layer, obtained by summing channels formed by applying an equivariant linear operator, a pointwise non-linearity, and either an invariant or equivariant linear output layer. Recently, Maron et al. (2019) showed that by allowing higher-order tensorization inside the network, universal invariant GNNs can be obtained. As a first contribution, we propose an alternative proof of this result, which relies on the Stone-Weierstrass theorem for algebra of real-valued functions. Our main contribution is then an extension of this result to the \emph{equivariant} case, which appears in many practical applications but has been less studied from a theoretical point of view. The proof relies on a new generalized Stone-Weierstrass theorem for algebra of equivariant functions, which is of

independent interest. Additionally, unlike many previous works that consider a fixed number of nodes, our results show that a GNN defined by a single set of parameters can approximate uniformly well a function defined on graphs of varying size.
****
A Primal-Dual link between GANs and Autoencoders
Hisham Husain, Richard Nock, 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.
****
Transfusion: Understanding Transfer Learning for Medical Imaging
Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, Samy Bengio
Transfer learning from natural image datasets, particularly ImageNet, using standard large models and corresponding pretrained weights has become a de-facto method for deep learning applications to medical imaging.
However, there are fundamental differences in data sizes, features and task specifications between natural image classification and the target medical tasks, and there is little understanding of the effects of transfer. In this paper, we explore properties of transfer learning for medical imaging. A performance evaluation on two large scale medical imaging tasks shows that surprisingly, transfer offers little benefit to performance, and simple, lightweight models can perform comparably to ImageNet architectures. Investigating the learned representations and features, we find that some of the differences from transfer learning are due to the over-parametrization of standard models rather than sophisticated feature reuse. We isolate where useful feature reuse occurs, and outline the implications for more efficient model exploration. We also explore feature independent benefits of transfer arising from weight scalings.
****
PIDForest: Anomaly Detection via Partial Identification
Parikshit Gopalan, Vatsal Sharan, Udi Wieder
We consider the problem of detecting anomalies in a large dataset. We propose a framework called Partial Identification which captures the intuition that anomalies are easy to distinguish from the overwhelming majority of points by relatively few attribute values. Formalizing this intuition, we propose a geometric anomaly measure for a point that we call PIDScore, which measures the minimum density of data points over all subcubes containing the point. We present PIDForest: a random forest based algorithm that finds anomalies based on this definition. We show that it performs favorably in comparison to several popular anomaly detection methods, across a broad range of benchmarks. PIDForest also provides a succinct explanation for why a point is labelled anomalous, by providing a set of features and ranges for them which are relatively uncommon in the dataset.
****
The Randomized Midpoint Method for Log-Concave Sampling
Ruoqi Shen, 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.
****
Face Reconstruction from Voice using Generative Adversarial Networks
Yandong Wen, Bhiksha Raj, Rita Singh
Voice profiling aims at inferring various human parameters from their speech, e.g. gender, age, etc. In this paper, we address the challenge posed by a subtask of voice profiling - reconstructing someone's face from their voice. The task is designed to answer the question: given an audio clip spoken by an unseen person, can we picture a face that has as many common elements, or associations as possible with the speaker, in terms of identity?
****

Using a Logarithmic Mapping to Enable Lower Discount Factors in Reinforcement Learning

Harm Van Seijen, Mehdi Fatemi, Arash Tavakoli

In an effort to better understand the different ways in which the discount factor affects the optimization process in reinforcement learning, we designed a set of experiments to study each effect in isolation. Our analysis reveals that the common perception that poor performance of low discount factors is caused by (too) small action-gaps requires revision. We propose an alternative hypothesis that identifies the size-difference of the action-gap across the state-space as the primary cause. We then introduce a new method that enables more homogeneous action-gaps by mapping value estimates to a logarithmic space. We prove convergence for this method under standard assumptions and demonstrate empirically that it indeed enables lower discount factors for approximate reinforcement-learning methods. This in turn allows tackling a class of reinforcement-learning problems that are challenging to solve with traditional methods.

****

PRNet: Self-Supervised Learning for Partial-to-Partial Registration

Yue Wang, Justin M. Solomon

We present a simple, flexible, and general framework titled Partial Registration Network (PRNet), for partial-to-partial point cloud registration. Inspired by recently-proposed learning-based methods for registration, we use deep networks to tackle non-convexity of the alignment and partial correspondence problem. While previous learning-based methods assume the entire shape is visible, PRNet is suitable for partial-to-partial registration, outperforming PointNetLK, DCP, and non-learning methods on synthetic data. PRNet is self-supervised, jointly learning an appropriate geometric representation, a keypoint detector that finds points in common between partial views, and keypoint-to-keypoint correspondences. We show PRNet predicts keypoints and correspondences consistently across views and objects. Furthermore, the learned representation is transferable to classification.

****

Adversarial Music: Real world Audio Adversary against Wake-word Detection System

Juncheng Li, Shuhui Qu, Xinjian Li, Joseph Szurley, J. Zico Kolter, Florian Metze

Voice Assistants (VAs) such as Amazon Alexa or Google Assistant rely on wake-word detection to respond to people's commands, which could potentially be vulnerable to audio adversarial examples. In this work, we target our attack on the wake-word detection system. Our goal is to jam the model with some inconspicuous background music to deactivate the VAs while our audio adversary is present. We implemented an emulated wake-word detection system of Amazon Alexa based on recent publications. We validated our models against the real Alexa in terms of wake-word detection accuracy. Then we computed our audio adversaries with consideration of expectation over transform and we implemented our audio adversary with a differentiable synthesizer. Next we verified our audio adversaries digitally on hundreds of samples of utterances collected from the real world. Our experiments show that we can effectively reduce the recognition F1 score of our emulated model from 93.4% to 11.0%. Finally, we tested our audio adversary over the air, and verified it works effectively against Alexa, reducing its F1 score from 92.5% to 11.0%. To the best of our knowledge, this is the first real-world adversarial attack against a commercial grade VA wake-word detection system. Our demo video is included in the supplementary material.

****

Learning to Optimize in Swarms

Yue Cao, Tianlong Chen, Zhangyang Wang, Yang Shen

Learning to optimize has emerged as a powerful framework for various optimization and machine learning tasks. Current such "meta-optimizers" often learn in the space of continuous optimization algorithms that are point-based and uncertainty-unaware. To overcome the limitations, we propose a meta-optimizer that learns in the algorithmic space of both point-based and population-based optimization algorithms. The meta-optimizer targets at a meta-loss function consisting of both

cumulative regret and entropy. Specifically, we learn and interpret the update f
ormula through a population of LSTMs embedded with sample- and feature-level att
entions. Meanwhile, we estimate the posterior directly over the global optimum a
nd use an uncertainty measure to help guide the learning process. Empirical resu
lts over non-convex test functions and the protein-docking application demonstra
te that this new meta-optimizer outperforms existing competitors. The codes are
publicly available at: https://github.com/Shen-Lab/LOIS
****
A Little Is Enough: Circumventing Defenses For Distributed Learning
Gilad Baruch, Moran Baruch, Yoav Goldberg
Distributed learning is central for large-scale training of deep-learning models
. However, it is exposed to a security threat in which Byzantine participants ca
n interrupt or control the learning process. Previous attack models assume that
the rogue participants (a) are omniscient (know the data of all other participan
ts), and (b) introduce large changes to the parameters.
Accordingly, most defense mechanisms make a similar assumption and attempt to us
e statistically robust methods to identify and discard values whose reported gra
dients are far from the population mean. We observe that if the empirical varian
ce between the gradients of workers is high enough, an attacker could take advan
tage of this and launch a non-omniscient attack that operates within the populat
ion variance. We show that the variance is indeed high enough even for simple da
tasets such as MNIST, allowing an attack that is not only undetected by existing
 defenses, but also uses their power against them, causing those defense mechani
sms to consistently select the byzantine workers while discarding legitimate one
s. We demonstrate our attack method works not only for preventing convergence bu
t also for repurposing of the model behavior (``backdooring''). We show that les
s than 25\% of colluding workers are sufficient to degrade the accuracy of mode
ls trained on MNIST, CIFAR10 and CIFAR100 by 50\%, as well as to introduce backd
oors without hurting the accuracy for MNIST and CIFAR10 datasets, but with a deg
radation for CIFAR100.
****
Statistical Model Aggregation via Parameter Matching
Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoan
g
We consider the problem of aggregating models learned from sequestered, possibly
 heterogeneous datasets. Exploiting tools from Bayesian nonparametrics, we devel
op a general meta-modeling framework that learns shared global latent structures
 by identifying correspondences among local model parameterizations. Our propose
d framework is model-independent and is applicable to a wide range of model type
s. After verifying our approach on simulated data, we demonstrate its utility in
 aggregating Gaussian topic models, hierarchical Dirichlet process based hidden
Markov models, and sparse Gaussian processes with applications spanning text sum
marization, motion capture analysis, and temperature forecasting.
****
Imitation Learning from Observations by Minimizing Inverse Dynamics Disagreement
Chao Yang, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Huaping Liu, Junzhou Huang, C
huang Gan
This paper studies Learning from Observations (LfO) for imitation learning with
access to state-only demonstrations. In contrast to Learning from Demonstration
(LfD) that involves both action and state supervisions, LfO is more practical in
 leveraging previously inapplicable resources (e.g., videos), yet more challengi
ng due to the incomplete expert guidance. In this paper, we investigate LfO and
its difference with LfD in both theoretical and practical perspectives. We first
 prove that the gap between LfD and LfO actually lies in the disagreement of inv
erse dynamics models between the imitator and expert, if following the modeling
approach of GAIL. More importantly, the upper bound of this gap is revealed by a
 negative causal entropy which can be minimized in a model-free way. We term our
 method as Inverse-Dynamics-Disagreement-Minimization (IDDM) which enhances the
conventional LfO method through further bridging the gap to LfD. Considerable em
pirical results on challenging benchmarks indicate that our method attains consi

stent improvements over other LfO counterparts.
****
Prediction of Spatial Point Processes: Regularized Method with Out-of-Sample Guarantees
Muhammad Osama, Dave Zachariah, Peter Stoica
A spatial point process can be characterized by an intensity function which predicts the number of events that occur across space. In this paper, we develop a method to infer predictive intensity intervals by learning a spatial model using a regularized criterion. We prove that the proposed method exhibits out-of-sample prediction performance guarantees which, unlike standard estimators, are valid even when the spatial model is misspecified. The method is demonstrated using synthetic as well as real spatial data.
****
STREETS: A Novel Camera Network Dataset for Traffic Flow
Corey Snyder, Minh Do
In this paper, we introduce STREETS, a novel traffic flow dataset from publicly available web cameras in the suburbs of Chicago, IL. We seek to address the limitations of existing datasets in this area. Many such datasets lack a coherent traffic network graph to describe the relationship between sensors. The datasets that do provide a graph depict traffic flow in urban population centers or highway systems and use costly sensors like induction loops. These contexts differ from that of a suburban traffic body. Our dataset provides over 4 million still images across 2.5 months and one hundred web cameras in suburban Lake County, IL. We divide the cameras into two distinct communities described by directed graphs and count vehicles to track traffic statistics. Our goal is to give researchers a benchmark dataset for exploring the capabilities of inexpensive and non-invasive sensors like web cameras to understand complex traffic bodies in communities of any size. We present benchmarking tasks and baseline results for one such task to guide how future work may use our dataset.
****
A Meta-Analysis of Overfitting in Machine Learning
Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, Ludwig Schmidt
We conduct the first large meta-analysis of overfitting due to test set reuse in the machine learning community. Our analysis is based on over one hundred machine learning competitions hosted on the Kaggle platform over the course of several years. In each competition, numerous practitioners repeatedly evaluated their progress against a holdout set that forms the basis of a public ranking available throughout the competition. Performance on a separate test set used only once determined the final ranking. By systematically comparing the public ranking with the final ranking, we assess how much participants adapted to the holdout set over the course of a competition. Our study shows, somewhat surprisingly, little evidence of substantial overfitting. These findings speak to the robustness of the holdout method across different data domains, loss functions, model classes, and human analysts.
****
Projected Stein Variational Newton: A Fast and Scalable Bayesian Inference Method in High Dimensions
Peng Chen, Keyi Wu, Joshua Chen, Tom O'Leary-Roseberry, Omar Ghattas
We propose a projected Stein variational Newton (pSVN) method for high-dimensional Bayesian inference. To address the curse of dimensionality, we exploit the intrinsic low-dimensional geometric structure of the posterior distribution in the high-dimensional parameter space via its Hessian (of the log posterior) operator and perform a parallel update of the parameter samples projected into a low-dimensional subspace by an SVN method. The subspace is adaptively constructed using the eigenvectors of the averaged Hessian at the current samples. We demonstrate fast convergence of the proposed method, complexity independent of the parameter and sample dimensions, and parallel scalability.
****
From deep learning to mechanistic understanding in neuroscience: the structure o

f retinal prediction
Hidenori Tanaka, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen Baccus, Surya Ganguli
Recently, deep feedforward neural networks have achieved considerable success in modeling biological sensory processing, in terms of reproducing the input-output map of sensory neurons. However, such models raise profound questions about the very nature of explanation in neuroscience. Are we simply replacing one complex system (a biological circuit) with another (a deep network), without understanding either? Moreover, beyond neural representations, are the deep network's computational mechanisms for generating neural responses the same as those in the brain? Without a systematic approach to extracting and understanding computational mechanisms from deep neural network models, it can be difficult both to assess the degree of utility of deep learning approaches in neuroscience, and to extract experimentally testable hypotheses from deep networks. We develop such a systematic approach by combining dimensionality reduction and modern attribution methods for determining the relative importance of interneurons for specific visual computations. We apply this approach to deep network models of the retina, revealing a conceptual understanding of how the retina acts as a predictive feature extractor that signals deviations from expectations for diverse spatiotemporal stimuli. For each stimulus, our extracted computational mechanisms are consistent with prior scientific literature, and in one case yields a new mechanistic hypothesis. Thus overall, this work not only yields insights into the computational mechanisms underlying the striking predictive capabilities of the retina, but also places the framework of deep networks as neuroscientific models on firmer theoretical foundations, by providing a new roadmap to go beyond comparing neural representations to extracting and understand computational mechanisms.
****
Abstract Reasoning with Distracting Features
Kecheng Zheng, Zheng-Jun Zha, Wei Wei
Abstraction reasoning is a long-standing challenge in artificial intelligence. Recent studies suggest that many of the deep architectures that have triumphed over other domains failed to work well in abstract reasoning. In this paper, we first illustrate that one of the main challenges in such a reasoning task is the presence of distracting features, which requires the learning algorithm to leverage counter-evidence and to reject any of false hypothesis in order to learn the true patterns. We later show that carefully designed learning trajectory over different categories of training data can effectively boost learning performance by mitigating the impacts of distracting features. Inspired this fact, we propose feature robust abstract reasoning (FRAR) model, which consists of a reinforcement learning based teacher network to determine the sequence of training and a student network for predictions. Experimental results demonstrated strong improvements over baseline algorithms and we are able to beat the state-of-the-art models by 18.7\% in RAVEN dataset and 13.3\% in the PGM dataset.
****
Deep Scale-spaces: Equivariance Over Scale
Daniel Worrall, Max Welling
We introduce deep scale-spaces, a generalization of convolutional neural networks, exploiting the scale symmetry structure of conventional image recognition tasks. Put plainly, the class of an image is invariant to the scale at which it is viewed. We construct scale equivariant cross-correlations based on a principled extension of convolutions, grounded in the theory of scale-spaces and semigroups. As a very basic operation, these cross-correlations can be used in almost any modern deep learning architecture in a plug-and-play manner. We demonstrate our networks on the Patch Camelyon and Cityscapes datasets, to prove their utility and perform introspective studies to further understand their properties.
****
Differentially Private Anonymized Histograms
Ananda Theertha Suresh
For a dataset of label-count pairs, an anonymized histogram is the multiset of counts. Anonymized histograms appear in various potentially sensitive contexts su

ch as password-frequency lists, degree distribution in social networks, and estimation of symmetric properties of discrete distributions. Motivated by these applications, we propose the first differentially private mechanism to release anonymized histograms that achieves near-optimal privacy utility trade-off both in terms of number of items and the privacy parameter. Further, if the underlying histogram is given in a compact format, the proposed algorithm runs in time sub-linear in the number of items. For anonymized histograms generated from unknown discrete distributions, we show that the released histogram can be directly used for estimating symmetric properties of the underlying distribution.
****
Generalized Sliced Wasserstein Distances
Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, Gustavo Rohde
The Wasserstein distance and its variations, e.g., the sliced-Wasserstein (SW) distance, have recently drawn attention from the machine learning community. The SW distance, specifically, was shown to have similar properties to the Wasserstein distance, while being much simpler to compute, and is therefore used in various applications including generative modeling and general supervised/unsupervised learning. In this paper, we first clarify the mathematical connection between the SW distance and the Radon transform. We then utilize the generalized Radon transform to define a new family of distances for probability measures, which we call generalized sliced-Wasserstein (GSW) distances. We further show that, similar to the SW distance, the GSW distance can be extended to a maximum GSW (max-GSW) distance. We then provide the conditions under which GSW and max-GSW distances are indeed proper metrics. Finally, we compare the numerical performance of the proposed distances on the generative modeling task of SW flows and report favorable results.
****
Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator
Arnak Dalalyan, Philip Thompson
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.
****
Scalable Spike Source Localization in Extracellular Recordings using Amortized Variational Inference
Cole Hurwitz, Kai Xu, Akash Srivastava, Alessio Buccino, Matthias Hennig
Determining the positions of neurons in an extracellular recording is useful for investigating the functional properties of the underlying neural circuitry. In this work, we present a Bayesian modelling approach for localizing the source of individual spikes on high-density, microelectrode arrays. To allow for scalable inference, we implement our model as a variational autoencoder and perform amortized variational inference. We evaluate our method on both biophysically realistic simulated and real extracellular datasets, demonstrating that it is more accurate than and can improve spike sorting performance over heuristic localization methods such as center of mass.
****
Prior-Free Dynamic Auctions with Low Regret Buyers
Yuan Deng, Jon Schneider, Balasubramanian Sivan
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.
****
When does label smoothing help?
Rafael Müller, Simon Kornblith, Geoffrey E. Hinton
The generalization and learning speed of a multi-class neural network can often be significantly improved by using soft targets that are a weighted average of the hard targets and the uniform distribution over labels. Smoothing the labels i

n this way prevents the network from becoming over-confident and label smoothing has been used in many state-of-the-art models, including image classification, language translation and speech recognition. Despite its widespread use, label smoothing is still poorly understood. Here we show empirically that in addition to improving generalization, label smoothing improves model calibration which can significantly improve beam search. However, we also observe that if a teacher network is trained with label smoothing, knowledge distillation into a student network is much less effective. To explain these observations, we visualize how label smoothing changes the representations learned by the penultimate layer of the network. We show that label smoothing encourages the representations of training examples from the same class to group in tight clusters. This results in loss of information in the logits about resemblances between instances of different classes, which is necessary for distillation, but does not hurt generalization or calibration of the model's predictions.
****
A General Framework for Symmetric Property Estimation
Moses Charikar, Kirankumar Shiragur, Aaron Sidford
In this paper we provide a general framework for estimating symmetric properties of distributions from i.i.d. samples. For a broad class of symmetric properties we identify the {\em easy} region where empirical estimation works and the {\em difficult} region where more complex estimators are required. We show that by approximately computing the profile maximum likelihood (PML) distribution \cite{ADOS16} in this difficult region we obtain a symmetric property estimation framework that is sample complexity optimal for many properties in a broader parameter regime than previous universal estimation approaches based on PML. The resulting algorithms based on these \emph{pseudo PML distributions} are also more practical.
****
Deep Generative Video Compression
Salvator Lombardo, JUN HAN, Christopher Schroers, Stephan Mandt
The usage of deep generative models for image compression has led to impressive performance gains over classical codecs while neural video compression is still in its infancy. Here, we propose an end-to-end, deep generative modeling approach to compress temporal sequences with a focus on video. Our approach builds upon variational autoencoder (VAE) models for sequential data and combines them with recent work on neural image compression. The approach jointly learns to transform the original sequence into a lower-dimensional representation as well as to discretize and entropy code this representation according to predictions of the sequential VAE. Rate-distortion evaluations on small videos from public data sets with varying complexity and diversity show that our model yields competitive results when trained on generic video content. Extreme compression performance is achieved when training the model on specialized content.
****
CondConv: Conditionally Parameterized Convolutions for Efficient Inference
Brandon Yang, Gabriel Bender, Quoc V. Le, Jiquan Ngiam
Convolutional layers are one of the basic building blocks of modern deep neural networks. One fundamental assumption is that convolutional kernels should be shared for all examples in a dataset. We propose conditionally parameterized convolutions (CondConv), which learn specialized convolutional kernels for each example. Replacing normal convolutions with CondConv enables us to increase the size and capacity of a network, while maintaining efficient inference. We demonstrate that scaling networks with CondConv improves the performance and inference cost trade-off of several existing convolutional neural network architectures on both classification and detection tasks. On ImageNet classification, our CondConv approach applied to EfficientNet-B0 achieves state-of the-art performance of 78.3% accuracy with only 413M multiply-adds. Code and checkpoints for the CondConv Tensorflow layer and CondConv-EfficientNet models are available at: https://github.com/tensorflow/tpu/tree/master/ models/official/efficientnet/condconv.
****

Towards a Zero-One Law for Column Subset Selection
Zhao Song, David Woodruff, Peilin Zhong
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 Attribution for Semantic Bug-Localization in Student Programs
Rahul Gupta, Aditya Kanade, Shirish Shevade
Providing feedback is an integral part of teaching. Most open online courses on programming make use of automated grading systems to support programming assignments and give real-time feedback. These systems usually rely on test results to quantify the programs' functional correctness. They return failing tests to the students as feedback. However, students may find it difficult to debug their programs if they receive no hints about where the bug is and how to fix it. In this work, we present NeuralBugLocator, a deep learning based technique, that can localize the bugs in a faulty program with respect to a failing test, without even running the program. At the heart of our technique is a novel tree convolutional neural network which is trained to predict whether a program passes or fails a given test. To localize the bugs, we analyze the trained network using a state-of-the-art neural prediction attribution technique and see which lines of the programs make it predict the test outcomes. Our experiments show that NeuralBugLocator is generally more accurate than two state-of-the-art program-spectrum based and one syntactic difference based bug-localization baselines.
****
Theoretical Limits of Pipeline Parallel Optimization and Application to Distributed Deep Learning
Igor Colin, Ludovic DOS SANTOS, Kevin Scaman
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.
****
DppNet: Approximating Determinantal Point Processes with Deep Networks
Zelda E. Mariet, Yaniv Ovadia, Jasper Snoek
Determinantal point processes (DPPs) provide an elegant and versatile way to sample sets of items that balance the point-wise quality with the set-wise diversity of selected items. For this reason, they have gained prominence in many machine learning applications that rely on subset selection. However, sampling from a DPP over a ground set of size N is a costly operation, requiring in general an O(N^3) preprocessing cost and an O(Nk^3) sampling cost for subsets of size k. We approach this problem by introducing DppNets: generative deep models that produce DPP-like samples for arbitrary ground sets. We develop an inhibitive attention mechanism based on transformer networks that captures a notion of dissimilarity between feature vectors. We show theoretically that such an approximation is sensible as it maintains the guarantees of inhibition or dissimilarity that makes DPPs so powerful and unique. Empirically, we show across multiple datasets that DPPNET is orders of magnitude faster than competing approaches for DPP sampling, while generating high-likelihood samples and performing as well as DPPs on downstream tasks.
****
Nonzero-sum Adversarial Hypothesis Testing Games
Sarath Yasodharan, Patrick Loiseau
We study nonzero-sum hypothesis testing games that arise in the context of adversarial classification, in both the Bayesian as well as the Neyman-Pearson frameworks. We first show that these games admit mixed strategy Nash equilibria, and then we examine some interesting concentration phenomena of these equilibria. Our main results are on the exponential rates of convergence of classification errors at equilibrium, which are analogous to the well-known Chernoff-Stein lemma and Chernoff information that describe the error exponents in the classical binary

hypothesis testing problem, but with parameters derived from the adversarial model. The results are validated through numerical experiments.
****
Global Sparse Momentum SGD for Pruning Very Deep Neural Networks
Xiaohan Ding, guiguang ding, Xiangxin Zhou, Yuchen Guo, Jungong Han, Ji Liu
Deep Neural Network (DNN) is powerful but computationally expensive and memory intensive, thus impeding its practical usage on resource-constrained front-end devices. DNN pruning is an approach for deep model compression, which aims at eliminating some parameters with tolerable performance degradation. In this paper, we propose a novel momentum-SGD-based optimization method to reduce the network complexity by on-the-fly pruning. Concretely, given a global compression ratio, we categorize all the parameters into two parts at each training iteration which are updated using different rules. In this way, we gradually zero out the redundant parameters, as we update them using only the ordinary weight decay but no gradients derived from the objective function. As a departure from prior methods that require heavy human works to tune the layer-wise sparsity ratios, prune by solving complicated non-differentiable problems or finetune the model after pruning, our method is characterized by 1) global compression that automatically finds the appropriate per-layer sparsity ratios; 2) end-to-end training; 3) no need for a time-consuming re-training process after pruning; and 4) superior capability to find better winning tickets which have won the initialization lottery.
****
Thompson Sampling and Approximate Inference
My Phan, Yasin Abbasi Yadkori, Justin Domke
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.
****
Quantum Wasserstein Generative Adversarial Networks
Shouvanik Chakrabarti, Huang Yiming, Tongyang Li, Soheil Feizi, Xiaodi Wu
The study of quantum generative models is well-motivated, not only because of its importance in quantum machine learning and quantum chemistry but also because of the perspective of its implementation on near-term quantum machines. Inspired by previous studies on the adversarial training of classical and quantum generative models, we propose the first design of quantum Wasserstein Generative Adversarial Networks (WGANs), which has been shown to improve the robustness and the scalability of the adversarial training of quantum generative models even on noisy quantum hardware. Specifically, we propose a definition of the Wasserstein semimetric between quantum data, which inherits a few key theoretical merits of its classical counterpart. We also demonstrate how to turn the quantum Wasserstein semimetric into a concrete design of quantum WGANs that can be efficiently implemented on quantum machines. Our numerical study, via classical simulation of quantum systems, shows the more robust and scalable numerical performance of our quantum WGANs over other quantum GAN proposals. As a surprising application, our quantum WGAN has been used to generate a 3-qubit quantum circuit of ~50 gates that well approximates a 3-qubit 1-d Hamiltonian simulation circuit that requires over 10k gates using standard techniques.
****
Deep Learning without Weight Transport
Mohamed Akrout, Collin Wilson, Peter Humphreys, Timothy Lillicrap, Douglas B. Tweed
Current algorithms for deep learning probably cannot run in the brain because they rely on weight transport, where forward-path neurons transmit their synaptic
weights to a feedback path, in a way that is likely impossible biologically. An algorithm called feedback alignment achieves deep learning without weight transport by using random feedback weights, but it performs poorly on hard visual-recognition tasks. Here we describe two mechanisms — a neural circuit called a weight mirror and a modification of an algorithm proposed by Kolen and Pollack in 199

4 – both of which let the feedback path learn appropriate synaptic weights quickly and accurately even in large networks, without weight transport or complex wiring. Tested on the ImageNet visual-recognition task, these mechanisms outperform both feedback alignment and the newer sign-symmetry method, and nearly match backprop, the standard algorithm of deep learning, which uses weight transport.
****
Implicit Regularization of Discrete Gradient Dynamics in Linear Neural Networks
Gauthier Gidel, Francis Bach, Simon Lacoste-Julien
When optimizing over-parameterized models, such as deep neural networks, a large set of parameters can achieve zero training error. In such cases, the choice of the optimization algorithm and its respective hyper-parameters introduces biases that will lead to convergence to specific minimizers of the objective. Consequently, this choice can be considered as an implicit regularization for the training of over-parametrized models. In this work, we push this idea further by studying the discrete gradient dynamics of the training of a two-layer linear network with the least-squares loss. Using a time rescaling, we show that, with a vanishing initialization and a small enough step size, this dynamics sequentially learns the solutions of a reduced-rank regression with a gradually increasing rank.
****
Generative Models for Graph-Based Protein Design
John Ingraham, Vikas Garg, Regina Barzilay, Tommi Jaakkola
Engineered proteins offer the potential to solve many problems in biomedicine, energy, and materials science, but creating designs that succeed is difficult in practice. A significant aspect of this challenge is the complex coupling between protein sequence and 3D structure, with the task of finding a viable design often referred to as the inverse protein folding problem. We develop relational language models for protein sequences that directly condition on a graph specification of the target structure. Our approach efficiently captures the complex dependencies in proteins by focusing on those that are long-range in sequence but local in 3D space. Our framework significantly improves in both speed and robustness over conventional and deep-learning-based methods for structure-based protein sequence design, and takes a step toward rapid and targeted biomolecular design with the aid of deep generative models.
****
Spike-Train Level Backpropagation for Training Deep Recurrent Spiking Neural Networks
Wenrui Zhang, Peng Li
Spiking neural networks (SNNs) well support spatiotemporal learning and energy-efficient event-driven hardware neuromorphic processors. As an important class of SNNs, recurrent spiking neural networks (RSNNs) possess great computational power. However, the practical application of RSNNs is severely limited by challenges in training. Biologically-inspired unsupervised learning has limited capability in boosting the performance of RSNNs. On the other hand, existing backpropagation (BP) methods suffer from high complexity of unrolling in time, vanishing and exploding gradients, and approximate differentiation of discontinuous spiking activities when applied to RSNNs. To enable supervised training of RSNNs under a well-defined loss function, we present a novel Spike-Train level RSNNs Backpropagation (ST-RSBP) algorithm for training deep RSNNs. The proposed ST-RSBP directly computes the gradient of a rated-coded loss function defined at the output layer of the network w.r.t tunable parameters. The scalability of ST-RSBP is achieved by the proposed spike-train level computation during which temporal effects of the SNN is captured in both the forward and backward pass of BP. Our ST-RSBP algorithm can be broadly applied to RSNNs with a single recurrent layer or deep RSNNs with multiple feed-forward and recurrent layers. Based upon challenging speech and image datasets including TI46, N-TIDIGITS, Fashion-MNIST and MNIST, ST-RSBP is able to train RSNNs with an accuracy surpassing that of the current state-of-art SNN BP algorithms and conventional non-spiking deep learning models.
****

Fully Parameterized Quantile Function for Distributional Reinforcement Learning
Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, Tie-Yan Liu
Distributional Reinforcement Learning (RL) differs from traditional RL in that, rather than the expectation of total returns, it estimates distributions and has achieved state-of-the-art performance on Atari Games. The key challenge in practical distributional RL algorithms lies in how to parameterize estimated distributions so as to better approximate the true continuous distribution. Existing distributional RL algorithms parameterize either the probability side or the return value side of the distribution function, leaving the other side uniformly fixed as in C51, QR-DQN or randomly sampled as in IQN. In this paper, we propose fully parameterized quantile function that parameterizes both the quantile fraction axis (i.e., the x-axis) and the value axis (i.e., y-axis) for distributional RL. Our algorithm contains a fraction proposal network that generates a discrete set of quantile fractions and a quantile value network that gives corresponding quantile values. The two networks are jointly trained to find the best approximation of the true distribution. Experiments on 55 Atari Games show that our algorithm significantly outperforms existing distributional RL algorithms and creates a new record for the Atari Learning Environment for non-distributed agents.
****
Neural Taskonomy: Inferring the Similarity of Task-Derived Representations from Brain Activity
Aria Wang, Michael Tarr, Leila Wehbe
Convolutional neural networks (CNNs) trained for object classification have been widely used to account for visually-driven neural responses in both human and primate brains. However, because of the generality and complexity of object classification, despite the effectiveness of CNNs in predicting brain activity, it is difficult to draw specific inferences about neural information processing using CNN-derived representations. To address this problem, we used learned representations drawn from 21 computer vision tasks to construct encoding models for predicting brain responses from BOLD5000---a large-scale dataset comprised of fMRI scans collected while observers viewed over 5000 naturalistic scene and object images. Encoding models based on task features predict activity in different regions across the whole brain. Features from 3D tasks such as keypoint/edge detection explain greater variance compared to 2D tasks---a pattern observed across the whole brain. Using results across all 21 task representations, we constructed a ``task graph'' based on the spatial layout of well-predicted brain areas from each task. A comparison of this brain-derived task structure to the task structure derived from transfer learning accuracy demonstrate that tasks with higher transferability make similar predictions for brain responses from different regions. These results---arising out of state-of-the-art computer vision methods---help reveal the task-specific architecture of the human visual system.
****
Adaptive Gradient-Based Meta-Learning Methods
Mikhail Khodak, Maria-Florina F. Balcan, Ameet S. Talwalkar
We build a theoretical framework for designing and understanding practical meta-learning methods that integrates sophisticated formalizations of task-similarity with the extensive literature on online convex optimization and sequential prediction algorithms. Our approach enables the task-similarity to be learned adaptively, provides sharper transfer-risk bounds in the setting of statistical learning-to-learn, and leads to straightforward derivations of average-case regret bounds for efficient algorithms in settings where the task-environment changes dynamically or the tasks share a certain geometric structure. We use our theory to modify several popular meta-learning algorithms and improve their training and meta-test-time performance on standard problems in few-shot and federated learning.
****
Compositional generalization through meta sequence-to-sequence learning
Brenden M. Lake
People can learn a new concept and use it compositionally, understanding how to "blicket twice" after learning how to "blicket." In contrast, powerful sequence-

to-sequence (seq2seq) neural networks fail such tests of compositionality, especially when composing new concepts together with existing concepts. In this paper, I show how memory-augmented neural networks can be trained to generalize compositionally through meta seq2seq learning. In this approach, models train on a series of seq2seq problems to acquire the compositional skills needed to solve new seq2seq problems. Meta se2seq learning solves several of the SCAN tests for compositional learning and can learn to apply implicit rules to variables.
****
Meta-Learning Representations for Continual Learning
Khurram Javed, Martha White
The reviews had two major concerns: lack of a benchmarking on a complex dataset, and unclear writing. To address these two major issues we:
1- Rewrote experiments section with improved terminology to make the paper more clear. Previously we were using the term Pretraining to refer to both a baseline and the meta-training stage. As the reviewers pointed out, this was confusing. We have replaced one of the usages with 'meta-training.' We have also changed evaluation to meta-testing.
2- Added mini-imagenet experiments to show that the proposed method scales to more complex datasets.
****
Massively scalable Sinkhorn distances via the Nyström method
Jason Altschuler, Francis Bach, Alessandro Rudi, Jonathan Niles-Weed
The Sinkhorn "distance," a variant of the Wasserstein distance with entropic regularization, is an increasingly popular tool in machine learning and statistical inference. However, the time and memory requirements of standard algorithms for computing this distance grow quadratically with the size of the data, rendering them prohibitively expensive on massive data sets. In this work, we show that this challenge is surprisingly easy to circumvent: combining two simple techniques—the Nyström method and Sinkhorn scaling—provably yields an accurate approximation of the Sinkhorn distance with significantly lower time and memory requirements than other approaches. We prove our results via new, explicit analyses of the Nyström method and of the stability properties of Sinkhorn scaling. We validate our claims experimentally by showing that our approach easily computes Sinkhorn distances on data sets hundreds of times larger than can be handled by other techniques.
****
Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling
Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, Qibin Zhao
Tensor-based multimodal fusion techniques have exhibited great predictive performance. However, one limitation is that existing approaches only consider bilinear or trilinear pooling, which fails to unleash the complete expressive power of multilinear fusion with restricted orders of interactions. More importantly, simply fusing features all at once ignores the complex local intercorrelations, leading to the deterioration of prediction. In this work, we first propose a polynomial tensor pooling (PTP) block for integrating multimodal features by considering high-order moments, followed by a tensorized fully connected layer. Treating PTP as a building block, we further establish a hierarchical polynomial fusion network (HPFN) to recursively transmit local correlations into global ones. By stacking multiple PTPs, the expressivity capacity of HPFN enjoys an exponential growth w.r.t. the number of layers, which is shown by the equivalence to a very deep convolutional arithmetic circuits. Various experiments demonstrate that it can achieve the state-of-the-art performance.
****
A Composable Specification Language for Reinforcement Learning Tasks
Kishor Jothimurugan, Rajeev Alur, Osbert Bastani
Reinforcement learning is a promising approach for learning control policies for robot tasks. However, specifying complex tasks (e.g., with multiple objectives and safety constraints) can be challenging, since the user must design a reward function that encodes the entire task. Furthermore, the user often needs to manually shape the reward to ensure convergence of the learning algorithm. We propos

e a language for specifying complex control tasks, along with an algorithm that compiles specifications in our language into a reward function and automatically performs reward shaping. We implement our approach in a tool called SPECTRL, and show that it outperforms several state-of-the-art baselines.
****

Learning to Predict 3D Objects with an Interpolation-based Differentiable Renderer

Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, Sanja Fidler

Many machine learning models operate on images, but ignore the fact that images are 2D projections formed by 3D geometry interacting with light, in a process called rendering. Enabling ML models to understand image formation might be key for generalization. However, due to an essential rasterization step involving discrete assignment operations, rendering pipelines are non-differentiable and thus largely inaccessible to gradient-based ML techniques. In this paper, we present DIB-Render, a novel rendering framework through which gradients can be analytically computed. Key to our approach is to view rasterization as a weighted interpolation, allowing image gradients to back-propagate through various standard vertex shaders within a single framework. Our approach supports optimizing over vertex positions, colors, normals, light directions and texture coordinates, and allows us to incorporate various well-known lighting models from graphics. We showcase our approach in two ML applications: single-image 3D object prediction, and 3D textured object generation, both trained using exclusively 2D supervision.
****

On the Utility of Learning about Humans for Human-AI Coordination

Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, Anca Dragan

While we would like agents that can coordinate with humans, current algorithms such as self-play and population-based training create agents that can coordinate with themselves. Agents that assume their partner to be optimal or similar to them can converge to coordination protocols that fail to understand and be understood by humans. To demonstrate this, we introduce a simple environment that requires challenging coordination, based on the popular game Overcooked, and learn a simple model that mimics human play. We evaluate the performance of agents trained via self-play and population-based training. These agents perform very well when paired with themselves, but when paired with our human model, they are significantly worse than agents designed to play with the human model. An experiment with a planning algorithm yields the same conclusion, though only when the human-aware planner is given the exact human model that it is playing with. A user study with real humans shows this pattern as well, though less strongly. Qualitatively, we find that the gains come from having the agent adapt to the human's gameplay. Given this result, we suggest several approaches for designing agents that learn about humans in order to better coordinate with them. Code is available at https://github.com/HumanCompatibleAI/overcooked_ai.
****

FastSpeech: Fast, Robust and Controllable Text to Speech

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu

Neural network based end-to-end text to speech (TTS) has significantly improved the quality of synthesized speech. Prominent methods (e.g., Tacotron 2) usually first generate mel-spectrogram from text, and then synthesize speech from the mel-spectrogram using vocoder such as WaveNet. Compared with traditional concatenative and statistical parametric approaches, neural network based end-to-end models suffer from slow inference speed, and the synthesized speech is usually not robust (i.e., some words are skipped or repeated) and lack of controllability (voice speed or prosody control). In this work, we propose a novel feed-forward network based on Transformer to generate mel-spectrogram in parallel for TTS. Specifically, we extract attention alignments from an encoder-decoder based teacher model for phoneme duration prediction, which is used by a length regulator to expand the source phoneme sequence to match the length of the target mel-spectrogram sequence for parallel mel-spectrogram generation. Experiments on the LJSpeech

dataset show that our parallel model matches autoregressive models in terms of speech quality, nearly eliminates the problem of word skipping and repeating in particularly hard cases, and can adjust voice speed smoothly. Most importantly, compared with autoregressive Transformer TTS, our model speeds up mel-spectrogram generation by 270x and the end-to-end speech synthesis by 38x. Therefore, we call our model FastSpeech.
****

Maximum Expected Hitting Cost of a Markov Decision Process and Informativeness of Rewards
Falcon Dai, Matthew Walter
We propose a new complexity measure for Markov decision processes (MDPs), the maximum expected hitting cost (MEHC). This measure tightens the closely related notion of diameter [JOA10] by accounting for the reward structure.
We show that this parameter replaces diameter in the upper bound on the optimal value span of an extended MDP, thus refining the associated upper bounds on the regret of several UCRL2-like algorithms.
Furthermore, we show that potential-based reward shaping [NHR99] can induce equivalent reward functions with varying informativeness, as measured by MEHC.
By analyzing the change in the maximum expected hitting cost, this work presents a formal understanding of the effect of potential-based reward shaping on regret (and sample complexity) in the undiscounted average reward setting.
We further establish that shaping can reduce or increase MEHC by at most a factor of two in a large class of MDPs with finite MEHC and unsaturated optimal average rewards.
****

Park: An Open Platform for Learning-Augmented Computer Systems
Hongzi Mao, Parimarjan Negi, Akshay Narayan, Hanrui Wang, Jiacheng Yang, Haonan Wang, Ryan Marcus, ravichandra addanki, Mehrdad Khani Shirkoohi, Songtao He, Vikram Nathan, Frank Cangialosi, Shaileshh Venkatakrishnan, Wei-Hung Weng, Song Han, Tim Kraska, Dr.Mohammad Alizadeh
We present Park, a platform for researchers to experiment with Reinforcement Learning (RL) for computer systems. Using RL for improving the performance of systems has a lot of potential, but is also in many ways very different from, for example, using RL for games. Thus, in this work we first discuss the unique challenges RL for systems has, and then propose Park an open extensible platform, which makes it easier for ML researchers to work on systems problems. Currently, Park consists of 12 real world system-centric optimization problems with one common easy to use interface. Finally, we present the performance of existing RL approaches over those 12 problems and outline potential areas of future work.
****

Adaptive Influence Maximization with Myopic Feedback
Binghui Peng, Wei Chen
We study the adaptive influence maximization problem with myopic feedback under the independent cascade model: one sequentially selects k nodes as seeds one by one from a social network, and each selected seed returns the immediate neighbors it activates as the feedback available for by later selections, and the goal is to maximize the expected number of total activated nodes, referred as the influence spread. We show that the adaptivity gap, the ratio between the optimal adaptive influence spread and the optimal non-adaptive influence spread, is at most 4 and at least e/(e-1), and the approximation ratios with respect to the optimal adaptive influence spread of both the non-adaptive greedy and adaptive greedy algorithms are at least \frac{1}{4}(1 - \frac{1}{e}) and at most \frac{e^2 + 1}{(e + 1)^2} < 1 - \frac{1}{e}. Moreover, the approximation ratio of the non-adaptive greedy algorithm is no worse than that of the adaptive greedy algorithm, when considering all graphs.
Our result confirms a long-standing open conjecture of Golovin and Krause (2011) on the constant approximation ratio of adaptive greedy with myopic feedback, and it also suggests that adaptive greedy may not bring much benefit under myopic feedback.
****

Compression with Flows via Local Bits-Back Coding
Jonathan Ho, Evan Lohn, Pieter Abbeel
Likelihood-based generative models are the backbones of lossless compression due
 to the guaranteed existence of codes with lengths close to negative log likelih
ood. However, there is no guaranteed existence of computationally efficient code
s that achieve these lengths, and coding algorithms must be hand-tailored to spe
cific types of generative models to ensure computational efficiency. Such coding
 algorithms are known for autoregressive models and variational autoencoders, bu
t not for general types of flow models. To fill in this gap, we introduce local
bits-back coding, a new compression technique for flow models. We present effici
ent algorithms that instantiate our technique for many popular types of flows, a
nd we demonstrate that our algorithms closely achieve theoretical codelengths fo
r state-of-the-art flow models on high-dimensional data.
****
On Adversarial Mixup Resynthesis
Christopher Beckham, Sina Honari, Vikas Verma, Alex M. Lamb, Farnoosh Ghadiri, R
 Devon Hjelm, Yoshua Bengio, Chris Pal
In this paper, we explore new approaches to combining information encoded within
 the learned representations of auto-encoders. We explore models that are capabl
e of combining the attributes of multiple inputs such that a resynthesised outpu
t is trained to fool an adversarial discriminator for real versus synthesised da
ta. Furthermore, we explore the use of such an architecture in the context of se
mi-supervised learning, where we learn a mixing function whose objective is to p
roduce interpolations of hidden states, or masked combinations of latent represe
ntations that are consistent with a conditioned class label. We show quantitativ
e and qualitative evidence that such a formulation is an interesting avenue of r
esearch.
****
High Fidelity Video Prediction with Large Stochastic Recurrent Neural Networks
Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V. Le, Hongl
ak Lee
Predicting future video frames is extremely challenging, as there are many facto
rs of variation that make up the dynamics of how frames change through time. Pre
viously proposed solutions require complex inductive biases inside network archi
tectures with highly specialized computation, including segmentation masks, opti
cal flow, and foreground and background separation. In this work, we question if
 such handcrafted architectures are necessary and instead propose a different ap
proach: finding minimal inductive bias for video prediction while maximizing net
work capacity. We investigate this question by performing the first large-scale
empirical study and demonstrate state-of-the-art performance by learning large m
odels on three different datasets: one for modeling object interactions, one for
 modeling human motion, and one for modeling car driving.
****
Variational Bayes under Model Misspecification
Yixin Wang, David Blei
Variational Bayes (VB) is a scalable alternative to Markov chain Monte Carlo (MC
MC) for Bayesian posterior inference. Though popular, VB comes with few theoreti
cal guarantees, most of which focus on well-specified models. However, models ar
e rarely well-specified in practice. In this work, we study VB under model missp
ecification. We prove the VB posterior is asymptotically normal and centers at t
he value that minimizes the Kullback-Leibler (KL) divergence to the true data-ge
nerating distribution. Moreover, the VB posterior mean centers at the same value
 and is also asymptotically normal. These results generalize the variational Ber
nstein--von Mises theorem [29] to misspecified models. As a consequence of these
 results, we find that the model misspecification error dominates the variationa
l approximation error in VB posterior predictive distributions. It explains the
widely observed phenomenon that VB achieves comparable predictive accuracy with
MCMC even though VB uses an approximating family. As illustrations, we study VB
under three forms of model misspecification, ranging from model over-/under-disp
ersion to latent dimensionality misspecification. We conduct two simulation stud

ies that demonstrate the theoretical results.
****
Certifying Geometric Robustness of Neural Networks
Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, Martin Vechev
The use of neural networks in safety-critical computer vision systems calls for their
robustness certification against natural geometric transformations (e.g., rotation,
scaling). However, current certification methods target mostly norm-based pixel
perturbations and cannot certify robustness against geometric transformations. In
this work, we propose a new method to compute sound and asymptotically optimal
linear relaxations for any composition of transformations. Our method is based on
a novel combination of sampling and optimization. We implemented the method
in a system called DeepG and demonstrated that it certifies significantly more
complex geometric transformations than existing methods on both defended and
undefended networks while scaling to large architectures.
****
Constrained deep neural network architecture search for IoT devices accounting for hardware calibration
Florian Scheidegger, Luca Benini, Costas Bekas, A. Cristiano I. Malossi
Deep neural networks achieve outstanding results for challenging image classification tasks. However, the design of network topologies is a complex task, and the research community is conducting ongoing efforts to discover top-accuracy topologies, either manually or by employing expensive architecture searches. We propose a unique narrow-space architecture search that focuses on delivering low-cost and rapidly executing networks that respect strict memory and time requirements typical of Internet-of-Things (IoT) near-sensor computing platforms. Our approach provides solutions with classification latencies below 10~ms running on a low-cost device with 1~GB RAM and a peak performance of 5.6~GFLOPS. The narrow-space search of floating-point models improves the accuracy on CIFAR10 of an established IoT model from 70.64% to 74.87% within the same memory constraints. We further improve the accuracy to 82.07% by including 16-bit half types and obtain the highest accuracy of 83.45% by extending the search with model-optimized IEEE 754 reduced types. To the best of our knowledge, this is the first empirical demonstration of more than 3000 trained models that run with reduced precision and push the Pareto optimal front by a wide margin. Within a given memory constraint, accuracy is improved by more than 7% points for half and more than 1% points for the best individual model format.
****
MAVEN: Multi-Agent Variational Exploration
Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, Shimon Whiteson
Centralised training with decentralised execution is an important setting for cooperative deep multi-agent reinforcement learning due to communication constraints during execution and computational tractability in training. In this paper, we analyse value-based methods that are known to have superior performance in complex environments. We specifically focus on QMIX, the current state-of-the-art in this domain. We show that the representation constraints on the joint action-values introduced by QMIX and similar methods lead to provably poor exploration and suboptimality. Furthermore, we propose a novel approach called MAVEN that hybridises value and policy-based methods by introducing a latent space for hierarchical control. The value-based agents condition their behaviour on the shared latent variable controlled by a hierarchical policy. This allows MAVEN to achieve committed, temporally extended exploration, which is key to solving complex multi-agent tasks. Our experimental results show that MAVEN achieves significant performance improvements on the challenging SMAC domain.
****
The continuous Bernoulli: fixing a pervasive error in variational autoencoders
Gabriel Loaiza-Ganem, John P. Cunningham

Variational autoencoders (VAE) have quickly become a central tool in machine learning, applicable to a broad range of data types and latent variable models. By far the most common first step, taken by seminal papers and by core software libraries alike, is to model MNIST data using a deep network parameterizing a Bernoulli likelihood. This practice contains what appears to be and what is often set aside as a minor inconvenience: the pixel data is [0,1] valued, not {0,1} as supported by the Bernoulli likelihood. Here we show that, far from being a triviality or nuisance that is convenient to ignore, this error has profound importance to VAE, both qualitative and quantitative. We introduce and fully characterize a new [0,1]-supported, single parameter distribution: the continuous Bernoulli, which patches this pervasive bug in VAE. This distribution is not nitpicking; it produces meaningful performance improvements across a range of metrics and datasets, including sharper image samples, and suggests a broader class of performant VAE.

****

Propagating Uncertainty in Reinforcement Learning via Wasserstein Barycenters

Alberto Maria Metelli, Amarildo Likmeta, Marcello Restelli

How does the uncertainty of the value function propagate when performing temporal difference learning? In this paper, we address this question by proposing a Bayesian framework in which we employ approximate posterior distributions to model the uncertainty of the value function and Wasserstein barycenters to propagate it across state-action pairs. Leveraging on these tools, we present an algorithm, Wasserstein Q-Learning (WQL), starting in the tabular case and then, we show how it can be extended to deal with continuous domains. Furthermore, we prove that, under mild assumptions, a slight variation of WQL enjoys desirable theoretical properties in the tabular setting. Finally, we present an experimental campaign to show the effectiveness of WQL on finite problems, compared to several RL algorithms, some of which are specifically designed for exploration, along with some preliminary results on Atari games.

****

DFNets: Spectral CNNs for Graphs with Feedback-Looped Filters

W. O. K. Asiri Suranga Wijesinghe, Qing Wang

We propose a novel spectral convolutional neural network (CNN) model on graph structured data, namely Distributed Feedback-Looped Networks (DFNets). This model is incorporated with a robust class of spectral graph filters, called feedback-looped filters, to provide better localization on vertices, while still attaining fast convergence and linear memory requirements. Theoretically, feedback-looped filters can guarantee convergence w.r.t. a specified error bound, and be applied universally to any graph without knowing its structure. Furthermore, the propagation rule of this model can diversify features from the preceding layers to produce strong gradient flows. We have evaluated our model using two benchmark tasks: semi-supervised document classification on citation networks and semi-supervised entity classification on a knowledge graph. The experimental results show that our model considerably outperforms the state-of-the-art methods in both benchmark tasks over all datasets.

****

Multiclass Learning from Contradictions

Sauptik Dhar, Vladimir Cherkassky, Mohak Shah

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-relational Poincaré Graph Embeddings

Ivana Balazevic, Carl Allen, Timothy Hospedales

Hyperbolic embeddings have recently gained attention in machine learning due to their ability to represent hierarchical data more accurately and succinctly than their Euclidean analogues. However, multi-relational knowledge graphs often exhibit multiple simultaneous hierarchies, which current hyperbolic models do not capture. To address this, we propose a model that embeds multi-relational graph d

ata in the Poincaré ball model of hyperbolic space. Our Multi-Relational Poincar
é model (MuRP) learns relation-specific parameters to transform entity embedding
s by Möbius matrix-vector multiplication and Möbius addition. Experiments on the
 hierarchical WN18RR knowledge graph show that our Poincaré embeddings outperfor
m their Euclidean counterpart and existing embedding methods on the link predict
ion task, particularly at lower dimensionality.
****
Verified Uncertainty Calibration
Ananya Kumar, Percy S. Liang, Tengyu Ma
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.
****
Episodic Memory in Lifelong Language Learning
Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, Dani Yogatama
We introduce a lifelong language learning setup where a model needs to learn fro
m a stream of text examples without any dataset identifier. We propose an episod
ic memory model that performs sparse experience replay and local adaptation to m
itigate catastrophic forgetting in this setup. Experiments on text classificatio
n and question answering demonstrate the complementary benefits of sparse experi
ence replay and local adaptation to allow the model to continuously learn from n
ew datasets. We also show that the space complexity of the episodic memory modul
e can be reduced significantly (~50-90%) by randomly choosing which examples to
store in memory with a minimal decrease in performance. We consider an episodic
memory component as a crucial building block of general linguistic intelligence
and see our model as a first step in that direction.
****
MetaQuant: Learning to Quantize by Learning to Penetrate Non-differentiable Quan
tization
Shangyu Chen, Wenya Wang, Sinno Jialin Pan
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.
****
Normalization Helps Training of Quantized LSTM
Lu Hou, Jinhua Zhu, James Kwok, Fei Gao, Tao Qin, Tie-Yan Liu
The long-short-term memory (LSTM), though powerful, is memory and computa\x02tio
n expensive. To alleviate this problem, one approach is to compress its weights
by quantization. However, existing quantization methods usually have inferior pe
rformance when used on LSTMs. In this paper, we first show theoretically that tr
aining a quantized LSTM is difficult because quantization makes the exploding gr
adient problem more severe, particularly when the LSTM weight matrices are large
. We then show that the popularly used weight/layer/batch normalization schemes
can help stabilize the gradient magnitude in training quantized LSTMs. Empirical
 results show that the normalized quantized LSTMs achieve significantly better r
esults than their unnormalized counterparts. Their performance is also comparabl
e with the full-precision LSTM, while being much smaller in size.
****
Differentially Private Bayesian Linear Regression
Garrett Bernstein, Daniel R. Sheldon
Linear regression is an important tool across many fields that work with sensiti
ve human-sourced data. Significant prior work has focused on producing different
ially private point estimates, which provide a privacy guarantee to individuals
while still allowing modelers to draw insights from data by estimating regressio
n coefficients. We investigate the problem of Bayesian linear regression, with t
he goal of computing posterior distributions that correctly quantify uncertainty
 given privately released statistics. We show that a naive approach that ignores
 the noise injected by the privacy mechanism does a poor job in realistic data s

ettings. We then develop noise-aware methods that perform inference over the pri vacy mechanism and produce correct posteriors across a wide range of scenarios.
****

Wasserstein Dependency Measure for Representation Learning
Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, Pi erre Sermanet
Mutual information maximization has emerged as a powerful learning objective for unsupervised representation learning obtaining state-of-the-art performance in applications such as object recognition, speech recognition, and reinforcement l earning. However, such approaches are fundamentally limited since a tight lower bound on mutual information requires sample size exponential in the mutual infor mation. This limits the applicability of these approaches for prediction tasks w ith high mutual information, such as in video understanding or reinforcement lea rning. In these settings, such techniques are prone to overfit, both in theory a nd in practice, and capture only a few of the relevant factors of variation. Thi s leads to incomplete representations that are not optimal for downstream tasks. In this work, we empirically demonstrate that mutual information-based represen tation learning approaches do fail to learn complete representations on a number of designed and real-world tasks. To mitigate these problems we introduce the W asserstein dependency measure, which learns more complete representations by usi ng the Wasserstein distance instead of the KL divergence in the mutual informati on estimator. We show that a practical approximation to this theoretically motiv ated solution, constructed using Lipschitz constraint techniques from the GAN li terature, achieves substantially improved results on tasks where incomplete repr esentations are a major challenge.
****

Multi-Agent Common Knowledge Reinforcement Learning
Christian Schroeder de Witt, Jakob Foerster, Gregory Farquhar, Philip Torr, Wend elin Boehmer, Shimon Whiteson
Cooperative multi-agent reinforcement learning often requires decentralised poli cies, which severely limit the agents' ability to coordinate their behaviour. In this paper, we show that common knowledge between agents allows for complex dec entralised coordination. Common knowledge arises naturally in a large number of decentralised cooperative multi-agent tasks, for example, when agents can recons truct parts of each others' observations. Since agents can independently agree o n their common knowledge, they can execute complex coordinated policies that con dition on this knowledge in a fully decentralised fashion. We propose multi-agen t common knowledge reinforcement learning (MACKRL), a novel stochastic actor-cri tic algorithm that learns a hierarchical policy tree. Higher levels in the hiera rchy coordinate groups of agents by conditioning on their common knowledge, or d elegate to lower levels with smaller subgroups but potentially richer common kno wledge. The entire policy tree can be executed in a fully decentralised fashion. As the lowest policy tree level consists of independent policies for each agent , MACKRL reduces to independently learnt decentralised policies as a special cas e. We demonstrate that our method can exploit common knowledge for superior per formance on complex decentralised coordination tasks, including a stochastic mat rix game and challenging problems in StarCraft II unit micromanagement.
****

Subspace Detours: Building Transport Plans that are Optimal on Subspace Projecti ons
Boris Muzellec, Marco Cuturi
Computing optimal transport (OT) between measures in high dimensions is doomed b y the curse of dimensionality. A popular approach to avoid this curse is to proj ect input measures on lower-dimensional subspaces (1D lines in the case of slice d Wasserstein distances), solve the OT problem between these reduced measures, a nd settle for the Wasserstein distance between these reductions, rather than tha t between the original measures. This approach is however difficult to extend to the case in which one wants to compute an OT map (a Monge map) between the orig inal measures. Since computations are carried out on lower-dimensional projectio ns, classical map estimation techniques can only produce maps operating in these

 reduced dimensions. We propose in this work two methods to extrapolate, from an
 transport map that is optimal on a subspace, one that is nearly optimal in the
entire space. We prove that the best optimal transport plan that takes such "sub
space detours" is a generalization of the Knothe-Rosenblatt transport. We show t
hat these plans can be explicitly formulated when comparing Gaussian measures (b
etween which the Wasserstein distance is commonly referred to as the Bures or Fr
échet distance). We provide an algorithm to select optimal subspaces given pairs
 of Gaussian measures, and study scenarios in which that mediating subspace can
be selected using prior information. We consider applications to semantic mediat
ion between elliptic word embeddings and domain adaptation with Gaussian mixture
 models.
****
The Broad Optimality of Profile Maximum Likelihood
Yi Hao, Alon Orlitsky
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.
****
Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers
Guang-He Lee, Yang Yuan, Shiyu Chang, Tommi Jaakkola
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.
****
Exact sampling of determinantal point processes with sublinear time preprocessin
g
Michal Derezinski, Daniele Calandriello, Michal Valko
We study the complexity of sampling from a distribution over all index subsets o
f the set {1, ..., n} with the probability of a subset S proportional to the det
erminant of the submatrix LS of some n x n positive semidefinite matrix L, where
 LS corresponds to the entries of L indexed by S. Known as a determinantal point
 process (DPP), this distribution is used in machine learning to induce diversit
y in subset selection. When sampling from DDPs, we often wish to sample multiple
 subsets S with small expected size k = E[|S|] << n from a very large matrix L,
so it is important to minimize the preprocessing cost of the procedure (performe
d once) as well as the sampling cost (performed repeatedly). For this purpose we
 provide DPP-VFX, a new algorithm which, given access only to L, samples exactly
 from a determinantal point process while satisfying the following two propertie
s: (1) its preprocessing cost is n poly(k), i.e., sublinear in the size of L, an
d (2) its sampling cost is poly(k), i.e., independent of the size of L. Prior to
 our results, state-of-the-art exact samplers required O(n^3) preprocessing time
 and sampling time linear in n or dependent on the spectral properties of L. We
furthermore give a reduction which allows using our algorithm for exact sampling
 from cardinality constrained determinantal point processes with n poly(k) time
preprocessing. Our implementation of DPP-VFX is provided at https://github.com/g
uilgautier/DPPy/.
****
Neural Diffusion Distance for Image Segmentation
Jian Sun, Zongben Xu
Diffusion distance is a spectral method for measuring distance among nodes on gr
aph considering global data structure. In this work, we propose a spec-diff-net
for computing diffusion distance on graph based on approximate spectral decompos
ition. The network is a differentiable deep architecture consisting of feature
extraction and diffusion distance modules for computing diffusion distance on im
age by end-to-end training. We design low resolution kernel matching loss and h
igh resolution segment matching loss to enforce the network's output to be consi
stent with human-labeled image segments. To compute high-resolution diffusion d
istance or segmentation mask, we design an up-sampling strategy by feature-atten

tional interpolation which can be learned when training spec-diff-net. With the learned diffusion distance, we propose a hierarchical image segmentation method outperforming previous segmentation methods. Moreover, a weakly supervised seman tic segmentation network is designed using diffusion distance and achieved promi sing results on PASCAL VOC 2012 segmentation dataset.
****
Experience Replay for Continual Learning
David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, Gregory Wayne
Interacting with a complex world involves continual learning, in which tasks and data distributions change over time. A continual learning system should demonst rate both plasticity (acquisition of new knowledge) and stability (preservation of old knowledge). Catastrophic forgetting is the failure of stability, in which new experience overwrites previous experience. In the brain, replay of past exp erience is widely believed to reduce forgetting, yet it has been largely overloo ked as a solution to forgetting in deep reinforcement learning. Here, we introdu ce CLEAR, a replay-based method that greatly reduces catastrophic forgetting in multi-task reinforcement learning. CLEAR leverages off-policy learning and behav ioral cloning from replay to enhance stability, as well as on-policy learning to preserve plasticity. We show that CLEAR performs better than state-of-the-art d eep learning techniques for mitigating forgetting, despite being significantly l ess complicated and not requiring any knowledge of the individual tasks being le arned.
****
Efficient online learning with kernels for adversarial large scale problems
Rémi Jézéquel, Pierre Gaillard, Alessandro Rudi
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.
****
KNG: The K-Norm Gradient Mechanism
Matthew Reimherr, Jordan Awan
This paper presents a new mechanism for producing sanitized statistical summarie s that achieve {\it differential privacy}, called the {\it K-Norm Gradient} Mech anism, or KNG. This new approach maintains the strong flexibility of the exponen tial mechanism, while achieving the powerful utility performance of objective pe rturbation. KNG starts with an inherent objective function (often an empirical r isk), and promotes summaries that are close to minimizing the objective by weigh ting according to how far the gradient of the objective function is from zero. Working with the gradient instead of the original objective function allows for additional flexibility as one can penalize using different norms. We show that, unlike the exponential mechanism, the noise added by KNG is asymptotically negl igible compared to the statistical error for many problems. In addition to theor etical guarantees on privacy and utility, we confirm the utility of KNG empirica lly in the settings of linear and quantile regression through simulations.
****
On the Downstream Performance of Compressed Word Embeddings
Avner May, Jian Zhang, Tri Dao, Christopher Ré
Compressing word embeddings is important for deploying NLP models in memory-cons trained settings. However, understanding what makes compressed embeddings perfor m well on downstream tasks is challenging---existing measures of compression qua lity often fail to distinguish between embeddings that perform well and those th at do not. We thus propose the eigenspace overlap score as a new measure. We rel ate the eigenspace overlap score to downstream performance by developing general ization bounds for the compressed embeddings in terms of this score, in the cont ext of linear and logistic regression. We then show that we can lower bound the eigenspace overlap score for a simple uniform quantization compression method, h elping to explain the strong empirical performance of this method. Finally, we s how that by using the eigenspace overlap score as a selection criterion between embeddings drawn from a representative set we compressed, we can efficiently ide

ntify the better performing embedding with up to 2x lower selection error rates than the next best measure of compression quality, and avoid the cost of training a separate model for each task of interest.
****
Primal-Dual Block Generalized Frank-Wolfe
Qi Lei, JIACHENG ZHUO, Constantine Caramanis, Inderjit S. Dhillon, Alexandros G. Dimakis
We propose a generalized variant of Frank-Wolfe algorithm for solving a class of sparse/low-rank optimization problems. Our formulation includes Elastic Net, regularized SVMs and phase retrieval as special cases. The proposed Primal-Dual Block Generalized Frank-Wolfe algorithm reduces the per-iteration cost while maintaining linear convergence rate.
The per iteration cost of our method depends on the structural complexity of the solution (i.e. sparsity/low-rank) instead of the ambient dimension.
We empirically show that our algorithm outperforms the state-of-the-art methods on (multi-class) classification tasks.
****
Nonparametric Density Estimation & Convergence Rates for GANs under Besov IPM Losses
Ananya Uppal, Shashank Singh, Barnabas Poczos
We study the problem of estimating a nonparametric probability distribution under a family of losses called Besov IPMs. This family is quite large, including, for example, L^p distances, total variation distance, and generalizations of both Wasserstein (earthmover's) and Kolmogorov-Smirnov distances. For a wide variety of settings, we provide both lower and upper bounds, identifying precisely how the choice of loss function and assumptions on the data distribution interact to determine the mini-max optimal convergence rate. We also show that, in many cases, linear distribution estimates, such as the empirical distribution or kernel density estimator, cannot converge at the optimal rate. These bounds generalize, unify, or improve on several recent and classical results. Moreover, IPMs can be used to formalize a statistical model of generative adversarial networks (GANs). Thus, we show how our results imply bounds on the statistical error of a GAN, showing, for example, that, in many cases, GANs can strictly outperform the best linear estimator.
****
Blended Matching Pursuit
Cyrille Combettes, Sebastian Pokutta
Matching pursuit algorithms are an important class of algorithms in signal processing and machine learning. We present a blended matching pursuit algorithm, combining coordinate descent-like steps with stronger gradient descent steps, for minimizing a smooth convex function over a linear space spanned by a set of atoms. We derive sublinear to linear convergence rates according to the smoothness and sharpness orders of the function and demonstrate computational superiority of our approach. In particular, we derive linear rates for a large class of non-strongly convex functions, and we demonstrate in experiments that our algorithm enjoys very fast rates of convergence and wall-clock speed while maintaining a sparsity of iterates very comparable to that of the (much slower) orthogonal matching pursuit.
****
Efficient Near-Optimal Testing of Community Changes in Balanced Stochastic Block Models
Aditya Gangrade, Praveen Venkatesh, Bobak Nazer, Venkatesh Saligrama
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.
****
Who is Afraid of Big Bad Minima? Analysis of gradient-flow in spiked matrix-tensor models
Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Lenka

Zdeborová
Gradient-based algorithms are effective for many machine learning tasks, but des pite ample recent effort and some progress, it often remains unclear why they wo rk in practice in optimising high-dimensional non-convex functions and why they find good minima instead of being trapped in spurious ones.Here we present a qua ntitative theory explaining this behaviour in a spiked matrix-tensor model.Our f ramework is based on the Kac-Rice analysis of stationary points and a closed-for m analysis of gradient-flow originating from statistical physics. We show that there is a well defined region of parameters where the gradient-flow algorithm f inds a good global minimum despite the presence of exponentially many spurious l ocal minima.
We show that this is achieved by surfing on saddles that have strong negative di rection towards the global minima, a phenomenon that is connected to a BBP-type threshold in the Hessian describing the critical points of the landscapes.
****
Online Convex Matrix Factorization with Representative Regions
Jianhao Peng, Olgica Milenkovic, Abhishek Agarwal
Matrix factorization (MF) is a versatile learning method that has found wide app lications in various data-driven disciplines. Still, many MF algorithms do not a dequately scale with the size of available datasets and/or lack interpretability . To improve the computational efficiency of the method, an online (streaming) M F algorithm was proposed in Mairal et al., 2010. To enable data interpretability , a constrained version of MF, termed convex MF, was introduced in Ding et al., 2010. In the latter work, the basis vectors are required to lie in the convex hu ll of the data samples, thereby ensuring that every basis can be interpreted as a weighted combination of data samples. No current algorithmic solutions for onl ine convex MF are known as it is challenging to find adequate convex bases witho ut having access to the complete dataset. We address both problems by proposing the first online convex MF algorithm that maintains a collection of constant-siz e sets of representative data samples needed for interpreting each of the basis (Ding et al., 2010) and has the same almost sure convergence guarantees as the o nline learning algorithm of Mairal et al., 2010. Our proof techniques combine ra ndom coordinate descent algorithms with specialized quasi-martingale convergence analysis. Experiments on synthetic and real world datasets show significant com putational savings of the proposed online convex MF method compared to classical convex MF. Since the proposed method maintains small representative sets of dat a samples needed for convex interpretations, it is related to a body of work in theoretical computer science, pertaining to generating point sets (Blum et al., 2016), and in computer vision, pertaining to archetypal analysis (Mei et al., 20 18). Nevertheless, it differs from these lines of work both in terms of the obje ctive and algorithmic implementations.
****
Differential Privacy Has Disparate Impact on Model Accuracy
Eugene Bagdasaryan, Omid Poursaeed, Vitaly Shmatikov
Differential privacy (DP) is a popular mechanism for training machine learning models with bounded leakage about the presence of specific points in the training data. The cost of differential privacy is a reduction in the model's accuracy. We demonstrate that in the neural networks trained using differentially private stochastic gradient descent (DP-SGD), this cost is not borne equally: accuracy of DP models drops much more for the underrepresented classes and subgroups.
****
Fair Algorithms for Clustering
Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, Maryam Negahbani
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.
****
The Cells Out of Sample (COOS) dataset and benchmarks for measuring out-of-sampl

e generalization of image classifiers
Alex Lu, Amy Lu, Wiebke Schormann, Marzyeh Ghassemi, David Andrews, Alan Moses
Understanding if classifiers generalize to out-of-sample datasets is a central problem in machine learning. Microscopy images provide a standardized way to measure the generalization capacity of image classifiers, as we can image the same classes of objects under increasingly divergent, but controlled factors of variation. We created a public dataset of 132,209 images of mouse cells, COOS-7 (Cells Out Of Sample 7-Class). COOS-7 provides a classification setting where four test datasets have increasing degrees of covariate shift: some images are random subsets of the training data, while others are from experiments reproduced months later and imaged by different instruments. We benchmarked a range of classification models using different representations, including transferred neural network features, end-to-end classification with a supervised deep CNN, and features from a self-supervised CNN. While most classifiers perform well on test datasets similar to the training dataset, all classifiers failed to generalize their performance to datasets with greater covariate shifts. These baselines highlight the challenges of covariate shifts in image data, and establish metrics for improving the generalization capacity of image classifiers.
****
Counting the Optimal Solutions in Graphical Models
Radu Marinescu, Rina Dechter
We introduce #opt, a new inference task for graphical models which calls for counting the number of optimal solutions of the model. We describe a novel variable elimination based approach for solving this task, as well as a depth-first branch and bound algorithm that traverses the AND/OR search space of the model. The key feature of the proposed algorithms is that their complexity is exponential in the induced width of the model only. It does not depend on the actual number of optimal solutions. Our empirical evaluation on various benchmarks demonstrates the effectiveness of the proposed algorithms compared with existing depth-first and best-first search based approaches that enumerate explicitly the optimal solutions.
****
Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems
Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, Rosalind Picard
Building an open-domain conversational agent is a challenging problem. Current evaluation methods, mostly post-hoc judgments of static conversation, do not capture conversation quality in a realistic interactive context. In this paper, we investigate interactive human evaluation and provide evidence for its necessity; we then introduce a novel, model-agnostic, and dataset-agnostic method to approximate it. In particular, we propose a self-play scenario where the dialog system talks to itself and we calculate a combination of proxies such as sentiment and semantic coherence on the conversation trajectory. We show that this metric is capable of capturing the human-rated quality of a dialog model better than any automated metric known to-date, achieving a significant Pearson correlation (r>.7, p<.05). To investigate the strengths of this novel metric and interactive evaluation in comparison to state-of-the-art metrics and human evaluation of static conversations, we perform extended experiments with a set of models, including several that make novel improvements to recent hierarchical dialog generation architectures through sentiment and semantic knowledge distillation on the utterance level. Finally, we open-source the interactive evaluation platform we built and the dataset we collected to allow researchers to efficiently deploy and evaluate dialog models.
****
Robust Multi-agent Counterfactual Prediction
Alexander Peysakhovich, Christian Kroer, Adam Lerer
We consider the problem of using logged data to make predictions about what would happen if we changed the `rules of the game' in a multi-agent system. This task is difficult because in many cases we observe actions individuals take but not

their private information or their full reward functions. In addition, agents are strategic, so when the rules change, they will also change their actions. Existing methods (e.g. structural estimation, inverse reinforcement learning) assume that agents' behavior comes from optimizing some utility or that the system is in equilibrium. They make counterfactual predictions by using observed actions to learn the underlying utility function (a.k.a. type) and then solving for the equilibrium of the counterfactual environment. This approach imposes heavy assumptions such as the rationality of the agents being observed and a correct model of the environment and agents' utility functions. We propose a method for analyzing the sensitivity of counterfactual conclusions to violations of these assumptions, which we call robust multi-agent counterfactual prediction (RMAC). We provide a first-order method for computing RMAC bounds. We apply RMAC to classic environments in market design: auctions, school choice, and social choice.
****

On Tractable Computation of Expected Predictions
Pasha Khosravi, YooJung Choi, Yitao Liang, Antonio Vergari, Guy Van den Broeck
Computing expected predictions of discriminative models is a fundamental task in machine learning that appears in many interesting applications such as fairness, handling missing values, and data analysis. Unfortunately, computing expectations of a discriminative model with respect to a probability distribution defined by an arbitrary generative model has been proven to be hard in general. In fact, the task is intractable even for simple models such as logistic regression and a naive Bayes distribution. In this paper, we identify a pair of generative and discriminative models that enables tractable computation of expectations, as well as moments of any order, of the latter with respect to the former in case of regression. Specifically, we consider expressive probabilistic circuits with certain structural constraints that support tractable probabilistic inference. Moreover, we exploit the tractable computation of high-order moments to derive an algorithm to approximate the expectations for classification scenarios in which exact computations are intractable. Our framework to compute expected predictions allows for handling of missing data during prediction time in a principled and accurate way and enables reasoning about the behavior of discriminative models. We empirically show our algorithm to consistently outperform standard imputation techniques on a variety of datasets. Finally, we illustrate how our framework can be used for exploratory data analysis.
****

Stagewise Training Accelerates Convergence of Testing Error Over SGD
Zhuoning Yuan, Yan Yan, Rong Jin, Tianbao Yang
Stagewise training strategy is widely used for learning neural networks, which runs a stochastic algorithm (e.g., SGD) starting with a relatively large step size (aka learning rate) and geometrically decreasing the step size after a number of iterations. It has been observed that the stagewise SGD has much faster convergence than the vanilla SGD with a polynomially decaying step size in terms of both training error and testing error. {\it But how to explain this phenomenon has been largely ignored by existing studies.} This paper provides some theoretical evidence for explaining this faster convergence. In particular, we consider a stagewise training strategy for minimizing empirical risk that satisfies the Polyak-\L ojasiewicz (PL) condition, which has been observed/proved for neural networks and also holds for a broad family of convex functions. For convex loss functions and two classes of ``nice-behaviored" non-convex objectives that are close to a convex function, we establish faster convergence of stagewise training than the vanilla SGD under the PL condition on both training error and testing error. Experiments on stagewise learning of deep residual networks exhibits that it satisfies one type of non-convexity assumption and therefore can be explained by our theory.
****

Specific and Shared Causal Relation Modeling and Mechanism-Based Clustering
Biwei Huang, Kun Zhang, Pengtao Xie, Mingming Gong, Eric P. Xing, Clark Glymour
State-of-the-art approaches to causal discovery usually assume a fixed underlying causal model. However, it is often the case that causal models vary across dom

ains or subjects, due to possibly omitted factors that affect the quantitative causal effects. As a typical example, causal connectivity in the brain network has been reported to vary across individuals, with significant differences across groups of people, such as autistics and typical controls. In this paper, we develop a unified framework for causal discovery and mechanism-based group identification. In particular, we propose a specific and shared causal model (SSCM), which takes into account the variabilities of causal relations across individuals/groups and leverages their commonalities to achieve statistically reliable estimation. The learned SSCM gives the specific causal knowledge for each individual as well as the general trend over the population. In addition, the estimated model directly provides the group information of each individual. Experimental results on synthetic and real-world data demonstrate the efficacy of the proposed method.
****

Computational Separations between Sampling and Optimization
Kunal Talwar
Two commonly arising computational tasks in Bayesian learning are Optimization (Maximum A Posteriori estimation) and Sampling (from the posterior distribution). In the convex case these two problems are efficiently reducible to each other. Recent work (Ma et al. 2019) shows that in the non-convex case, sampling can sometimes be provably faster. We present a simpler and stronger separation.
We then compare sampling and optimization in more detail and show that they are provably incomparable: there are families of continuous functions for which optimization is easy but sampling is NP-hard, and vice versa. Further, we show function families that exhibit a sharp phase transition in the computational complexity of sampling, as one varies the natural temperature parameter. Our results draw on a connection to analogous separations in the discrete setting which are well-studied.
****

Classification Accuracy Score for Conditional Generative Models
Suman Ravuri, Oriol Vinyals
Deep generative models (DGMs) of images are now sufficiently mature that they produce nearly photorealistic samples and obtain scores similar to the data distribution on heuristics such as Frechet Inception Distance (FID). These results, especially on large-scale datasets such as ImageNet, suggest that DGMs are learning the data distribution in a perceptually meaningful space and can be used in downstream tasks. To test this latter hypothesis, we use class-conditional generative models from a number of model classes—variational autoencoders, autoregressive models, and generative adversarial networks (GANs)—to infer the class labels of real data. We perform this inference by training an image classifier using only synthetic data and using the classifier to predict labels on real data. The performance on this task, which we call Classification Accuracy Score (CAS), reveals some surprising results not identified by traditional metrics and constitute our contributions. First, when using a state-of-the-art GAN (BigGAN-deep), Top-1 and Top-5 accuracy decrease by 27.9% and 41.6%, respectively, compared to the original data; and conditional generative models from other model classes, such as Vector-Quantized Variational Autoencoder-2 (VQ-VAE-2) and Hierarchical Autoregressive Models (HAMs), substantially outperform GANs on this benchmark. Second, CAS automatically surfaces particular classes for which generative models failed to capture the data distribution, and were previously unknown in the literature. Third, we find traditional GAN metrics such as Inception Score (IS) and FID neither predictive of CAS nor useful when evaluating non-GAN models. Furthermore, in order to facilitate better diagnoses of generative models, we open-source the proposed metric.
****

Unsupervised Meta-Learning for Few-Shot Image Classification
Siavash Khodadadeh, Ladislau Boloni, Mubarak Shah
Few-shot or one-shot learning of classifiers requires a significant inductive bias towards the type of task to be learned. One way to acquire this is by meta-learning on tasks similar to the target task. In this paper, we propose UMTRA, an

algorithm that performs unsupervised, model-agnostic meta-learning for classification tasks.

The meta-learning step of UMTRA is performed on a flat collection of unlabeled images. While we assume that these images can be grouped into a diverse set of classes and are relevant to the target task, no explicit information about the classes or any labels are needed. UMTRA uses random sampling and augmentation to create synthetic training tasks for meta-learning phase. Labels are only needed at the final target task learning step, and they can be as little as one sample per class.

On the Omniglot and Mini-Imagenet few-shot learning benchmarks, UMTRA outperforms every tested approach based on unsupervised learning of representations, while alternating for the best performance with the recent CACTUs algorithm. Compared to supervised model-agnostic meta-learning approaches, UMTRA trades off some classification accuracy for a reduction in the required labels of several orders of magnitude.

****

Transferable Normalization: Towards Improving Transferability of Deep Neural Networks

Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, Michael I. Jordan

Deep neural networks (DNNs) excel at learning representations when trained on large-scale datasets. Pre-trained DNNs also show strong transferability when fine-tuned to other labeled datasets. However, such transferability becomes weak when the target dataset is fully unlabeled as in Unsupervised Domain Adaptation (UDA). We envision that the loss of transferability may stem from the intrinsic limitation of the architecture design of DNNs. In this paper, we delve into the components of DNN architectures and propose Transferable Normalization (TransNorm) in place of existing normalization techniques. TransNorm is an end-to-end trainable layer to make DNNs more transferable across domains. As a general method, TransNorm can be easily applied to various deep neural networks and domain adaption methods, without introducing any extra hyper-parameters or learnable parameters. Empirical results justify that TransNorm not only improves classification accuracies but also accelerates convergence for mainstream DNN-based domain adaptation methods.

****

Semi-Implicit Graph Variational Auto-Encoders

Arman Hasanzadeh, Ehsan Hajiramezanali, Krishna Narayanan, Nick Duffield, Mingyuan Zhou, Xiaoning Qian

Semi-implicit graph variational auto-encoder (SIG-VAE) is proposed to expand the flexibility of variational graph auto-encoders (VGAE) to model graph data. SIG-VAE employs a hierarchical variational framework to enable neighboring node sharing for better generative modeling of graph dependency structure, together with a Bernoulli-Poisson link decoder. Not only does this hierarchical construction provide a more flexible generative graph model to better capture real-world graph properties, but also does SIG-VAE naturally lead to semi-implicit hierarchical variational inference that allows faithful modeling of implicit posteriors of given graph data, which may exhibit heavy tails, multiple modes, skewness, and rich dependency structures. SIG-VAE integrates a carefully designed generative model, well suited to model real-world sparse graphs, and a sophisticated variational inference network, which propagates the graph structural information and distribution uncertainty to capture complex posteriors. SIG-VAE clearly outperforms a simple combination of VGAE with variational inference, including semi-implicit variational inference~(SIVI) or normalizing flow (NF), which does not propagate uncertainty in its inference network, and provides more interpretable latent representations than VGAE does. Extensive experiments with a variety of graph data show that SIG-VAE significantly outperforms state-of-the-art methods on several different graph analytic tasks.

****

Efficient Approximation of Deep ReLU Networks for Functions on Low Dimensional Manifolds

Minshuo Chen, Haoming Jiang, Wenjing Liao, Tuo Zhao

Deep neural networks have revolutionized many real world applications, due to their flexibility in data fitting and accurate predictions for unseen data. A line of research reveals that neural networks can approximate certain classes of functions with an arbitrary accuracy, while the size of the network scales exponentially with respect to the data dimension. Empirical results, however, suggest that networks of moderate size already yield appealing performance. To explain such a gap, a common belief is that many data sets exhibit low dimensional structures, and can be modeled as samples near a low dimensional manifold. In this paper, we prove that neural networks can efficiently approximate functions supported on low dimensional manifolds. The network size scales exponentially in the approximation error, with an exponent depending on the intrinsic dimension of the data and the smoothness of the function. Our result shows that exploiting low dimensional data structures can greatly enhance the efficiency in function approximation by neural networks. We also implement a sub-network that assigns input data to their corresponding local neighborhoods, which may be of independent interest.

****

GOT: An Optimal Transport framework for Graph comparison

Hermina Petric Maretic, Mireille El Gheche, Giovanni Chierchia, Pascal Frossard

We present a novel framework based on optimal transport for the challenging problem of comparing graphs. Specifically, we exploit the probabilistic distribution of smooth graph signals defined with respect to the graph topology. This allows us to derive an explicit expression of the Wasserstein distance between graph signal distributions in terms of the graph Laplacian matrices. This leads to a structurally meaningful measure for comparing graphs, which is able to take into account the global structure of graphs, while most other measures merely observe local changes independently. Our measure is then used for formulating a new graph alignment problem, whose objective is to estimate the permutation that minimizes the distance between two graphs. We further propose an efficient stochastic algorithm based on Bayesian exploration to accommodate for the non-convexity of the graph alignment problem. We finally demonstrate the performance of our novel framework on different tasks like graph alignment, graph classification and graph signal prediction, and we show that our method leads to significant improvement with respect to the-state-of-art algorithms.

****

Multivariate Distributionally Robust Convex Regression under Absolute Error Loss

Jose Blanchet, Peter W. Glynn, Jun Yan, Zhengqing Zhou

This paper proposes a novel non-parametric multidimensional convex regression estimator which is designed to be robust to adversarial perturbations in the empirical measure. We minimize over convex functions the maximum (over Wasserstein perturbations of the empirical measure) of the absolute regression errors. The inner maximization is solved in closed form resulting in a regularization penalty involves the norm of the gradient. We show consistency of our estimator and a rate of convergence of order $\widetilde{O}\left(n^{-1/d}\right)$, matching the bounds of alternative estimators based on square-loss minimization. Contrary to all of the existing results, our convergence rates hold without imposing compactness on the underlying domain and with no a priori bounds on the underlying convex function or its gradient norm.

****

A Benchmark for Interpretability Methods in Deep Neural Networks

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, Been Kim

We propose an empirical measure of the approximate accuracy of feature importance estimates in deep neural networks. Our results across several large-scale image classification datasets show that many popular interpretability methods produce estimates of feature importance that are not better than a random designation of feature importance. Only certain ensemble based approaches---VarGrad and SmoothGrad-Squared---outperform such a random assignment of importance. The manner of ensembling remains critical, we show that some approaches do no better then the underlying method but carry a far higher computational burden.

****
Biases for Emergent Communication in Multi-agent Reinforcement Learning
Tom Eccles, Yoram Bachrach, Guy Lever, Angeliki Lazaridou, Thore Graepel
We study the problem of emergent communication, in which language arises because speakers and listeners must communicate information in order to solve tasks. In temporally extended reinforcement learning domains, it has proved hard to learn such communication without centralized training of agents, due in part to a difficult joint exploration problem. We introduce inductive biases for positive signalling and positive listening, which ease this problem. In a simple one-step environment, we demonstrate how these biases ease the learning problem. We also apply our methods to a more extended environment, showing that agents with these inductive biases achieve better performance, and analyse the resulting communications protocols.
****
Zero-shot Knowledge Transfer via Adversarial Belief Matching
Paul Micaelli, Amos J. Storkey
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.
****
Uniform Error Bounds for Gaussian Process Regression with Application to Safe Control
Armin Lederer, Jonas Umlauft, Sandra Hirche
Data-driven models are subject to model errors due to limited and noisy training data. Key to the application of such models in safety-critical domains is the quantification of their model error. Gaussian processes provide such a measure and uniform error bounds have been derived, which allow safe control based on these models. However, existing error bounds require restrictive assumptions. In this paper, we employ the Gaussian process distribution and continuity arguments to derive a novel uniform error bound under weaker assumptions. Furthermore, we demonstrate how this distribution can be used to derive probabilistic Lipschitz constants and analyze the asymptotic behavior of our bound. Finally, we derive safety conditions for the control of unknown dynamical systems based on Gaussian process models and evaluate them in simulations of a robotic manipulator.
****
Leader Stochastic Gradient Descent for Distributed Training of Deep Learning Models
Yunfei Teng, Wenbo Gao, François Chalus, Anna E. Choromanska, Donald Goldfarb, Adrian Weller
We consider distributed optimization under communication constraints for training deep learning models. We propose a new algorithm, whose parameter updates rely on two forces: a regular gradient step, and a corrective direction dictated by the currently best-performing worker (leader). Our method differs from the parameter-averaging scheme EASGD in a number of ways: (i) our objective formulation does not change the location of stationary points compared to the original optimization problem; (ii) we avoid convergence decelerations caused by pulling local workers descending to different local minima to each other (i.e. to the average of their parameters); (iii) our update by design breaks the curse of symmetry (the phenomenon of being trapped in poorly generalizing sub-optimal solutions in symmetric non-convex landscapes); and (iv) our approach is more communication efficient since it broadcasts only parameters of the leader rather than all workers. We provide theoretical analysis of the batch version of the proposed algorithm, which we call Leader Gradient Descent (LGD), and its stochastic variant (LSGD). Finally, we implement an asynchronous version of our algorithm and extend it to the multi-leader setting, where we form groups of workers, each represented by its own local leader (the best performer in a group), and update each worker with a corrective direction comprised of two attractive forces: one to the local, and one to the global leader (the best performer among all workers). The multi-leader setting is well-aligned with current hardware architecture, where local wo

rkers forming a group lie within a single computational node and different groups correspond to different nodes. For training convolutional neural networks, we empirically demonstrate that our approach compares favorably to state-of-the-art baselines.
****
Random deep neural networks are biased towards simple functions
Giacomo De Palma, Bobak Kiani, Seth Lloyd
We prove that the binary classifiers of bit strings generated by random wide deep neural networks with ReLU activation function are biased towards simple functions. The simplicity is captured by the following two properties. For any given input bit string, the average Hamming distance of the closest input bit string with a different classification is at least sqrt(n / (2π log n)), where n is the length of the string. Moreover, if the bits of the initial string are flipped randomly, the average number of flips required to change the classification grows linearly with n. These results are confirmed by numerical experiments on deep neural networks with two hidden layers, and settle the conjecture stating that random deep neural networks are biased towards simple functions. This conjecture was proposed and numerically explored in [Valle Pérez et al., ICLR 2019] to explain the unreasonably good generalization properties of deep learning algorithms. The probability distribution of the functions generated by random deep neural networks is a good choice for the prior probability distribution in the PAC-Bayesian generalization bounds. Our results constitute a fundamental step forward in the characterization of this distribution, therefore contributing to the understanding of the generalization properties of deep learning algorithms.
****
Discrete Object Generation with Reversible Inductive Construction
Ari Seff, Wenda Zhou, Farhan Damani, Abigail Doyle, Ryan P. Adams
The success of generative modeling in continuous domains has led to a surge of interest in generating discrete data such as molecules, source code, and graphs. However, construction histories for these discrete objects are typically not unique and so generative models must reason about intractably large spaces in order to learn.
Additionally, structured discrete domains are often characterized by strict constraints on what constitutes a valid object and generative models must respect these requirements in order to produce useful novel samples.
Here, we present a generative model for discrete objects employing a Markov chain where transitions are restricted to a set of local operations that preserve validity.
Building off of generative interpretations of denoising autoencoders, the Markov chain alternates between producing 1) a sequence of corrupted objects that are valid but not from the data distribution, and 2) a learned reconstruction distribution that attempts to fix the corruptions while also preserving validity.
This approach constrains the generative model to only produce valid objects, requires the learner to only discover local modifications to the objects, and avoids marginalization over an unknown and potentially large space of construction histories.
We evaluate the proposed approach on two highly structured discrete domains, molecules and Laman graphs, and find that it compares favorably to alternative methods at capturing distributional statistics for a host of semantically relevant metrics.
****
Adaptively Aligned Image Captioning via Adaptive Attention Time
Lun Huang, Wenmin Wang, Yaxian Xia, Jie Chen
Recent neural models for image captioning usually employ an encoder-decoder framework with an attention mechanism. However, the attention mechanism in such a framework aligns one single (attended) image feature vector to one caption word, assuming one-to-one mapping from source image regions and target caption words, which is never possible. In this paper, we propose a novel attention model, namely Adaptive Attention Time (AAT), to align the source and the target adaptively for image captioning. AAT allows the framework to learn how many attention steps

to take to output a caption word at each decoding step. With AAT, an image region can be mapped to an arbitrary number of caption words while a caption word can also attend to an arbitrary number of image regions. AAT is deterministic and differentiable, and doesn't introduce any noise to the parameter gradients. In this paper, we empirically show that AAT improves over state-of-the-art methods on the task of image captioning. Code is available at https://github.com/husthuaan/AAT.
****
Fully Dynamic Consistent Facility Location
Vincent Cohen-Addad, Niklas Oskar D. Hjuler, Nikos Parotsidis, David Saulpic, Chris Schwiegelshohn
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 Rematerialization for Deep Networks
Ravi Kumar, Manish Purohit, Zoya Svitkina, Erik Vee, Joshua Wang
When training complex neural networks, memory usage can be an important bottleneck. The question of when to rematerialize, i.e., to recompute intermediate values rather than retaining them in memory, becomes critical to achieving the best time and space efficiency. In this work we consider the rematerialization problem and devise efficient algorithms that use structural characterizations of computation graphs---treewidth and pathwidth---to obtain provably efficient rematerialization schedules. Our experiments demonstrate the performance of these algorithms on many common deep learning models.
****
Flow-based Image-to-Image Translation with Feature Disentanglement
Ruho Kondo, Keisuke Kawano, Satoshi Koide, Takuro Kutsuna
Learning non-deterministic dynamics and intrinsic factors from images obtained through physical experiments is at the intersection of machine learning and material science. Disentangling the origins of uncertainties involved in microstructure growth, for example, is of great interest because future states vary due to thermal fluctuation and other environmental factors. To this end we propose a flow-based image-to-image model, called Flow U-Net with Squeeze modules (FUNS), that allows us to disentangle the features while retaining the ability to generate highquality diverse images from condition images. Our model successfully captures probabilistic phenomena by incorporating a U-Net-like architecture into the flowbased model. In addition, our model automatically separates the diversity of target images into condition-dependent/independent parts. We demonstrate that the quality and diversity of the images generated for microstructure growth and CelebA datasets outperform existing variational generative models.
****