A graph similarity for deep learning
Seongmin Ok
Graph neural networks (GNNs) have been successful in learning representations fr
om graphs. Many popular GNNs follow the pattern of aggregate-transform: they agg
regate the neighbors' attributes and then transform the results of aggregation w
ith a learnable function. Analyses of these GNNs explain which pairs of non-iden
tical graphs have different representations. However, we still lack an understan
ding of how similar these representations will be. We adopt kernel distance and
propose transform-sum-cat as an alternative to aggregate-transform to reflect th
e continuous similarity between the node neighborhoods in the neighborhood aggre
gation. The idea leads to a simple and efficient graph similarity, which we name
 Weisfeiler-Leman similarity (WLS). In contrast to existing graph kernels, WLS i
s easy to implement with common deep learning frameworks. In graph classificatio
n experiments, transform-sum-cat significantly outperforms other neighborhood ag
gregation methods from popular GNN models. We also develop a simple and fast GNN
 model based on transform-sum-cat, which obtains, in comparison with widely used
 GNN models, (1) a higher accuracy in node classification, (2) a lower absolute
error in graph regression, and (3) greater stability in adversarial training of
graph generation.
**********

An Unsupervised Information-Theoretic Perceptual Quality Metric
Sangnie Bhardwaj, Ian Fischer, Johannes Ballé, Troy Chinen
Tractable models of human perception have proved to be challenging to build. Han
d-designed models such as MS-SSIM remain popular predictors of human image quali
ty judgements due to their simplicity and speed. Recent modern deep learning app
roaches can perform better, but they rely on supervised data which can be costly
 to gather: large sets of class labels such as ImageNet, image quality ratings,
or both. We combine recent advances in information-theoretic objective functions
 with a computational architecture informed by the physiology of the human visua
l system and unsupervised training on pairs of video frames, yielding our Percep
tual Information Metric (PIM). We show that PIM is competitive with supervised m
etrics on the recent and challenging BAPPS image quality assessment dataset and
outperforms them in predicting the ranking of image compression methods in CLIC
2020. We also perform qualitative experiments using the ImageNet-C dataset, and
establish that PIM is robust with respect to architectural details.
**********

Self-Supervised MultiModal Versatile Networks
Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovi■, Ja
son Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, Andrew Zisserman
Videos are a rich source of multi-modal supervision. In this work, we learn repr
esentations using self-supervision by leveraging three modalities naturally pres
ent in videos: visual, audio and language streams.
To this end, we introduce the notion of a multimodal versatile network -- a netw
ork that can ingest multiple modalities and whose representations enable downstr
eam tasks in multiple modalities. In particular, we explore how best to combine
the modalities, such that fine-grained representations of the visual and audio m
odalities can be maintained, whilst also integrating text into a common embeddin
g.
Driven by versatility, we also introduce a novel process of deflation, so that
the networks can be effortlessly applied to the visual data in the form of video
 or a static image.
We demonstrate how such networks trained on large collections of unlabelled vide
o data can be applied on video, video-text, image and audio tasks.
Equipped with these representations, we obtain state-of-the-art performance on m
ultiple challenging benchmarks including UCF101, HMDB51, Kinetics600, AudioSet a
nd ESC-50 when compared to previous self-supervised work. Our models are publicl
y available.
**********

Benchmarking Deep Inverse Models over time, and the Neural-Adjoint method
Simiao Ren, Willie Padilla, Jordan Malof

We consider the task of solving generic inverse problems, where one wishes to determine the hidden parameters of a natural system that will give rise to a particular set of measurements. Recently many new approaches based upon deep learning have arisen, generating promising results. We conceptualize these models as different schemes for efficiently, but randomly, exploring the space of possible inverse solutions. As a result, the accuracy of each approach should be evaluated as a function of time rather than a single estimated solution, as is often done now.
Using this metric, we compare several state-of-the-art inverse modeling approaches on four benchmark tasks: two existing tasks, a new 2-dimensional sinusoid task, and a challenging modern task of meta-material design. Finally, inspired by our conception of the inverse problem, we explore a simple solution that uses a deep neural network as a surrogate (i.e., approximation) for the forward model, and then uses backpropagation with respect to the model input to search for good inverse solutions. Variations of this approach - which we term the neural adjoint (NA) - have been explored recently on specific problems, and here we evaluate it comprehensively on our benchmark. We find that the addition of a simple novel loss term - which we term the boundary loss - dramatically improves the NA's performance, and it consequentially achieves the best (or nearly best) performance in all of our benchmark scenarios.
**********

Off-Policy Evaluation and Learning for External Validity under a Covariate Shift
Masatoshi Uehara, Masahiro Kato, Shota Yasui
We consider the evaluation and training of a new policy for the evaluation data by using the historical data obtained from a different policy. The goal of off-policy evaluation (OPE) is to estimate the expected reward of a new policy over the evaluation data, and that of off-policy learning (OPL) is to find a new policy that maximizes the expected reward over the evaluation data. Although the standard OPE and OPL assume the same distribution of covariate between the historical and evaluation data, there often exists a problem of a covariate shift,i.e., the distribution of the covariate of the historical data is different from that of the evaluation data. In this paper, we derive the efficiency bound of OPE under a covariate shift. Then, we propose doubly robust and efficient estimators for OPE and OPL under a covariate shift by using an estimator of the density ratio between the distributions of the historical and evaluation data. We also discuss other possible estimators and compare their theoretical properties. Finally, we confirm the effectiveness of the proposed estimators through experiments.
**********

Neural Methods for Point-wise Dependency Estimation
Yao-Hung Hubert Tsai, Han Zhao, Makoto Yamada, Louis-Philippe Morency, Russ R. Salakhutdinov
Since its inception, the neural estimation of mutual information (MI) has demonstrated the empirical success of modeling expected dependency between high-dimensional random variables. However, MI is an aggregate statistic and cannot be used to measure point-wise dependency between different events. In this work, instead of estimating the expected dependency, we focus on estimating point-wise dependency (PD), which quantitatively measures how likely two outcomes co-occur. We show that we can naturally obtain PD when we are optimizing MI neural variational bounds. However, optimizing these bounds is challenging due to its large variance in practice. To address this issue, we develop two methods (free of optimizing MI variational bounds): Probabilistic Classifier and Density-Ratio Fitting. We demonstrate the effectiveness of our approaches in 1) MI estimation, 2) self-supervised representation learning, and 3) cross-modal retrieval task.
**********

Fast and Flexible Temporal Point Processes with Triangular Maps
Oleksandr Shchur, Nicholas Gao, Marin Biloš, Stephan Günnemann
Temporal point process (TPP) models combined with recurrent neural networks prov ide a powerful framework for modeling continuous-time event data. While such mod els are flexible, they are inherently sequential and therefore cannot benefit fr om the parallelism of modern hardware. By exploiting the recent developments in the field of normalizing flows, we design TriTPP - a new class of non-recurrent TPP models, where both sampling and likelihood computation can be done in parall el. TriTPP matches the flexibility of RNN-based methods but permits several orde rs of magnitude faster sampling. This enables us to use the new model for variat ional inference in continuous-time discrete-state systems. We demonstrate the ad vantages of the proposed framework on synthetic and real-world datasets.
**********
Backpropagating Linearly Improves Transferability of Adversarial Examples
Yiwen Guo, Qizhang Li, Hao Chen
The vulnerability of deep neural networks (DNNs) to adversarial examples has dra wn great attention from the community. In this paper, we study the transferabili ty of such examples, which lays the foundation of many black-box attacks on DNNs . We revisit a not so new but definitely noteworthy hypothesis of Goodfellow et al.'s and disclose that the transferability can be enhanced by improving the lin earity of DNNs in an appropriate manner. We introduce linear backpropagation (Li nBP), a method that performs backpropagation in a more linear fashion using off- the-shelf attacks that exploit gradients. More specifically, it calculates forwa rd as normal but backpropagates loss as if some nonlinear activations are not en countered in the forward pass. Experimental results demonstrate that this simple yet effective method obviously outperforms current state-of-the-arts in craftin g transferable adversarial examples on CIFAR-10 and ImageNet, leading to more ef fective attacks on a variety of DNNs. Code at: https://github.com/qizhangli/linb p-attack.
**********
PyGlove: Symbolic Programming for Automated Machine Learning
Daiyi Peng, Xuanyi Dong, Esteban Real, Mingxing Tan, Yifeng Lu, Gabriel Bender, Hanxiao Liu, Adam Kraft, Chen Liang, Quoc Le
Neural networks are sensitive to hyper-parameter and architecture choices. Autom ated Machine Learning (AutoML) is a promising paradigm for automating these choi ces. Current ML software libraries, however, are quite limited in handling the d ynamic interactions among the components of AutoML. For example, efficient NAS a lgorithms, such as ENAS and DARTS, typically require an implementation coupling between the search space and search algorithm, the two key components in AutoML. Furthermore, implementing a complex search flow, such as searching architecture s within a loop of searching hardware configurations, is difficult. To summarize , changing the search space, search algorithm, or search flow in current ML libr aries usually requires a significant change in the program logic.
**********
Fourier Sparse Leverage Scores and Approximate Kernel Learning
Tamas Erdelyi, Cameron Musco, Christopher Musco
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Improved Algorithms for Online Submodular Maximization via First-order Regret Bo unds
Nicholas Harvey, Christopher Liaw, Tasuku Soma
We consider the problem of nonnegative submodular maximization in the online set ting. At time step t, an algorithm selects a set St $\in$ C $\subseteq$ 2^V where C is a feasi ble family of sets. An adversary then reveals a submodular function ft. The goal is to design an efficient algorithm for minimizing the expected approximate reg ret. In this work, we give a general approach for improving regret bounds in onl ine submodular maximization by exploiting “first-order” regret bounds for online

linear optimization.
- For monotone submodular maximization subject to a matroid, we give an efficient algorithm which achieves a $(1 - c/e - \varepsilon)$-regret of $O(\sqrt{kT \ln(n/k)})$ where $n$ is the size of the ground set, $k$ is the rank of the matroid, $\varepsilon > 0$ is a constant, and $c$ is the average curvature. Even without assuming any curvature (i.e., taking $c = 1$), this regret bound improves on previous results of Streeter et al. (2009) and Golovin et al. (2014).
- For nonmonotone, unconstrained submodular functions, we give an algorithm with 1/2-regret $O(\sqrt{nT})$, improving on the results of Roughgarden and Wang (2018). Our approach is based on Blackwell approachability; in particular, we give a novel first-order regret bound for the Blackwell instances that arise in this setting
**********

Synbols: Probing Learning Algorithms with Synthetic Datasets
Alexandre Lacoste, Pau Rodríguez López, Frederic Branchaud-Charron, Parmida Atighehchian, Massimo Caccia, Issam Hadj Laradji, Alexandre Drouin, Matthew Craddock, Laurent Charlin, David Vázquez
Progress in the field of machine learning has been fueled by the introduction of benchmark datasets pushing the limits of existing algorithms.
Enabling the design of datasets to test specific properties and failure modes of learning algorithms is thus a problem of high interest, as it has a direct impact on innovation in the field. In this sense, we introduce Synbols — Synthetic Symbols — a tool for rapidly generating new datasets with a rich composition of latent features rendered in low resolution images. Synbols leverages the large amount of symbols available in the Unicode standard and the wide range of artistic font provided by the open font community. Our tool's high-level interface provides a language for rapidly generating new distributions on the latent features, including various types of textures and occlusions. To showcase the versatility of Synbols, we use it to dissect the limitations and flaws in standard learning algorithms in various learning setups including supervised learning, active learning, out of distribution generalization, unsupervised representation learning, and object counting.
**********

Adversarially Robust Streaming Algorithms via Differential Privacy
Avinatan Hasidim, Haim Kaplan, Yishay Mansour, Yossi Matias, Uri Stemmer
A streaming algorithm is said to be adversarially robust if its accuracy guarantees are maintained even when the data stream is chosen maliciously, by an adaptive adversary. We establish a connection between adversarial robustness of streaming algorithms and the notion of differential privacy. This connection allows us to design new adversarially robust streaming algorithms that outperform the current state-of-the-art constructions for many interesting regimes of parameters.
**********

Trading Personalization for Accuracy: Data Debugging in Collaborative Filtering
Long Chen, Yuan Yao, Feng Xu, Miao Xu, Hanghang Tong
Collaborative filtering has been widely used in recommender systems. Existing work has primarily focused on improving the prediction accuracy mainly via either building refined models or incorporating additional side information, yet has largely ignored the inherent distribution of the input rating data.
In this paper, we propose a data debugging framework to identify overly personalized ratings whose existence degrades the performance of a given collaborative filtering model. The key idea of the proposed approach is to search for a small set of ratings whose editing (e.g., modification or deletion) would near-optimally improve the recommendation accuracy of a validation set. Experimental results demonstrate that the proposed approach can significantly improve the recommendation accuracy. Furthermore, we observe that the identified ratings significantly deviate from the average ratings of the corresponding items, and the proposed approach tends to modify them towards the average. This result sheds light on the design of future recommender systems in terms of balancing between the overall accuracy and personalization.
**********

Cascaded Text Generation with Markov Transformers

Yuntian Deng, Alexander Rush
The two dominant approaches to neural text generation are fully autoregressive m
odels, using serial beam search decoding, and non-autoregressive models, using p
arallel decoding with no output dependencies. This work proposes an autoregressi
ve model with sub-linear parallel time generation. Noting that conditional rando
m fields with bounded context can be decoded in parallel, we propose an efficien
t cascaded decoding approach for generating high-quality output. To parameterize
 this cascade, we introduce a Markov transformer, a variant of the popular fully
 autoregressive model that allows us to simultaneously decode with specific auto
regressive context cutoffs. This approach requires only a small modification fro
m standard autoregressive training, while showing competitive accuracy/speed tra
deoff compared to existing methods on five machine translation datasets.
**********

Improving Local Identifiability in Probabilistic Box Embeddings
Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, Andrew McCa
llum
Geometric embeddings have recently received attention for their natural ability
to represent transitive asymmetric relations via containment.  Box embeddings, w
here objects are represented by n-dimensional hyperrectangles, are a particularl
y promising example of such an embedding as they are closed under intersection a
nd their volume can be calculated easily, allowing them to naturally represent c
alibrated probability distributions.  The benefits of geometric embeddings also
introduce a problem of local identifiability, however, where whole neighborhoods
 of parameters result in equivalent loss which impedes learning. Prior work addr
essed some of these issues by using an approximation to Gaussian convolution ove
r the box parameters, however this intersection operation also increases the spa
rsity of the gradient.  In this work we model the box parameters with min and ma
x Gumbel distributions, which were chosen such that the space is still closed un
der the operation of intersection. The calculation of the expected intersection
volume involves all parameters, and we demonstrate experimentally that this dras
tically improves the ability of such models to learn.
**********

Permute-and-Flip: A new mechanism for differentially private selection
Ryan McKenna, Daniel R. Sheldon
We consider the problem of differentially private selection.  Given a finite set
 of candidate items, and a quality score for each item, our goal is to design a
differentially private mechanism that returns an item with a score that is as hi
gh as possible.  The most commonly used mechanism for this task is the exponenti
al mechanism.  In this work, we propose a new mechanism for this task based on a
 careful analysis of the privacy constraints. The expected score of our mechanis
m is always at least as large as the exponential mechanism, and can offer improv
ements up to a factor of two.  Our mechanism is simple to implement and runs  in
 linear time.
**********

Deep reconstruction of strange attractors from time series
William Gilpin
Experimental measurements of physical systems often have a limited number of ind
ependent channels, causing essential dynamical variables to remain unobserved. H
owever, many popular methods for unsupervised inference of latent dynamics from
experimental data implicitly assume that the measurements have higher intrinsic
dimensionality than the underlying system---making coordinate identification a d
imensionality reduction problem. Here, we study the opposite limit, in which hid
den governing coordinates must be inferred from only a low-dimensional time seri
es of measurements. Inspired by classical analysis techniques for partial observ
ations of chaotic attractors, we introduce a general embedding technique for uni
variate and multivariate time series, consisting of an autoencoder trained with
a novel latent-space loss function. We show that our technique reconstructs the
strange attractors of synthetic and real-world systems better than existing tech
niques, and that it creates consistent, predictive representations of even stoch
astic systems. We conclude by using our technique to discover dynamical attracto

rs in diverse systems such as patient electrocardiograms, household electricity usage, neural spiking, and eruptions of the Old Faithful geyser---demonstrating diverse applications of our technique for exploratory data analysis.
**********

Reciprocal Adversarial Learning via Characteristic Functions
Shengxi Li, Zeyang Yu, Min Xiang, Danilo Mandic
Generative adversarial nets (GANs) have become a preferred tool for tasks involving complicated distributions. To stabilise the training and reduce the mode collapse of GANs, one of their main variants employs the integral probability metric (IPM) as the loss function. This provides extensive IPM-GANs with theoretical support for basically comparing moments in an embedded domain of the \textit{critic}. We generalise this by comparing the distributions rather than their moments via a powerful tool, i.e., the characteristic function (CF), which uniquely and universally comprising all the information about a distribution. For rigour, we first establish the physical meaning of the phase and amplitude in CF, and show that this provides a feasible way of balancing the accuracy and diversity of generation. We then develop an efficient sampling strategy to calculate the CFs. Within this framework, we further prove an equivalence between the embedded and data domains when a reciprocal exists, where we naturally develop the GAN in an auto-encoder structure, in a way of comparing everything in the embedded space (a semantically meaningful manifold). This efficient structure uses only two modules, together with a simple training strategy, to achieve bi-directionally generating clear images, which is referred to as the reciprocal CF GAN (RCF-GAN). Experimental results demonstrate the superior performances of the proposed RCF-GAN in terms of both generation and reconstruction.
**********

Statistical Guarantees of Distributed Nearest Neighbor Classification
Jiexin Duan, Xingye Qiao, Guang Cheng
Nearest neighbor is a popular nonparametric method for classification and regression with many appealing properties. In the big data era, the sheer volume and spatial/temporal disparity of big data may prohibit centrally processing and storing the data. This has imposed considerable hurdle for nearest neighbor predictions since the entire training data must be memorized. One effective way to overcome this issue is the distributed learning framework. Through majority voting, the distributed nearest neighbor classifier achieves the same rate of convergence as its oracle version in terms of the regret, up to a multiplicative constant that depends solely on the data dimension. The multiplicative difference can be eliminated by replacing majority voting with the weighted voting scheme. In addition, we provide sharp theoretical upper bounds of the number of subsamples in order for the distributed nearest neighbor classifier to reach the optimal convergence rate. It is interesting to note that the weighted voting scheme allows a larger number of subsamples than the majority voting one. Our findings are supported by numerical studies.
**********

Stein Self-Repulsive Dynamics: Benefits From Past Samples
Mao Ye, Tongzheng Ren, Qiang Liu
We propose a new Stein self-repulsive dynamics for obtaining diversified samples from intractable un-normalized distributions. Our idea is to introduce Stein variational gradient as a repulsive force to push the samples of Langevin dynamics away from the past trajectories. This simple idea allows us to significantly decrease the auto-correlation in Langevin dynamics and hence increase the effective sample size. Importantly, as we establish in our theoretical analysis, the asymptotic stationary distribution remains correct even with the addition of the repulsive force,
thanks to the special properties of the Stein variational gradient. We perform extensive empirical studies of our new algorithm, showing that our method yields much higher sample efficiency and better uncertainty estimation than vanilla Langevin dynamics.
**********

The Statistical Complexity of Early-Stopped Mirror Descent

Tomas Vaskevicius, Varun Kanade, Patrick Rebeschini
Recently there has been a surge of interest in understanding implicit regulariza tion properties of iterative gradient-based optimization algorithms. In this pap er, we study the statistical guarantees on the excess risk achieved by early-sto pped unconstrained mirror descent algorithms applied to the unregularized empiri cal risk with the squared loss for linear models and kernel methods. By completi ng an inequality that characterizes convexity for the squared loss, we identify an intrinsic link between offset Rademacher complexities and potential-based con vergence analysis of mirror descent methods. Our observation immediately yields excess risk guarantees for the path traced by the iterates of mirror descent in terms of offset complexities of certain function classes depending only on the c hoice of the mirror map, initialization point, step-size, and the number of iter ations. We apply our theory to recover, in a rather clean and elegant manner via rather short proofs, some of the recent results in the implicit regularization literature, while also showing how to improve upon them in some settings.
**********

Algorithmic recourse under imperfect causal knowledge: a probabilistic approach
Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, Isabel Valera
Recent work has discussed the limitations of counterfactual explanations to reco mmend actions for algorithmic recourse, and argued for the need of taking causal relationships between features into consideration. Unfortunately, in practice, the true underlying structural causal model is generally unknown. In this work, we first show that it is impossible to guarantee recourse without access to the true structural equations. To address this limitation, we propose two probabilis tic approaches to select optimal actions that achieve recourse with high probabi lity given limited causal knowledge (e.g., only the causal graph). The first cap tures uncertainty over structural equations under additive Gaussian noise, and u ses Bayesian model averaging to estimate the counterfactual distribution. The se cond removes any assumptions on the structural equations by instead computing th e average effect of recourse actions on individuals similar to the person who se eks recourse, leading to a novel subpopulation-based interventional notion of re course. We then derive a gradient-based procedure for selecting optimal recourse actions, and empirically show that the proposed approaches lead to more reliabl e recommendations under imperfect causal knowledge than non-probabilistic baseli nes.
**********

Quantitative Propagation of Chaos for SGD in Wide Neural Networks
Valentin De Bortoli, Alain Durmus, Xavier Fontaine, Umut Simsekli
Requests for name changes in the electronic proceedings will be accepted with 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 Causal View on Robustness of Neural Networks
Cheng Zhang, Kun Zhang, Yingzhen Li
We present a causal view on the robustness of neural networks against input mani pulations, which applies not only to traditional classification tasks but also t o general measurement data. Based on this view, we design a deep causal manipula tion augmented model (deep CAMA) which explicitly models possible manipulations on certain causes leading to changes in the observed effect. We further develop data augmentation and test-time fine-tuning methods to improve deep CAMA's robu stness. When compared with discriminative deep neural networks, our proposed mod el shows superior robustness against unseen manipulations. As a by-product, our model achieves disentangled representation which separates the representation of manipulations from those of other latent causes.
**********

Minimax Classification with 0-1 Loss and Performance Guarantees
Santiago Mazuelas, Andrea Zanoni, Aritz Pérez
Supervised classification techniques use training samples to find classification rules with small expected 0-1 loss. Conventional methods achieve efficient lear

ning and out-of-sample generalization by minimizing surrogate losses over specif
ic families of rules. This paper presents minimax risk classifiers (MRCs) that d
o not rely on a choice of surrogate loss and family of rules. MRCs achieve effic
ient learning and out-of-sample generalization by minimizing worst-case expected
 0-1 loss w.r.t. uncertainty sets that are defined by linear constraints and inc
lude the true underlying distribution. In addition, MRCs' learning stage provide
s performance guarantees as lower and upper tight bounds for expected 0-1 loss.
We also present MRCs' finite-sample generalization bounds in terms of training s
ize and smallest minimax risk, and show their competitive classification perform
ance w.r.t. state-of-the-art techniques using benchmark datasets.
**********

How to Learn a Useful Critic? Model-based Action-Gradient-Estimator Policy Optim
ization

Pierluca D'Oro, Wojciech Ja■kowski

Deterministic-policy actor-critic algorithms for continuous control improve the
actor by plugging its actions into the critic and ascending the action-value gra
dient, which is obtained by chaining the actor's Jacobian matrix with the gradie
nt of the critic with respect to input actions. However, instead of gradients, t
he critic is, typically, only trained to accurately predict expected returns, wh
ich, on their own, are useless for policy optimization. In this paper, we propos
e MAGE, a model-based actor-critic algorithm, grounded in the theory of policy g
radients, which explicitly learns the action-value gradient. MAGE backpropagates
 through the learned dynamics to compute gradient targets in temporal difference
 learning, leading to a critic tailored for policy improvement. On a set of MuJo
Co continuous-control tasks, we demonstrate the efficiency of the algorithm in c
omparison to model-free and model-based state-of-the-art baselines.
**********

Coresets for Regressions with Panel Data

Lingxiao Huang, K Sudhir, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Learning Composable Energy Surrogates for PDE Order Reduction

Alex Beatson, Jordan Ash, Geoffrey Roeder, Tianju Xue, Ryan P. Adams

Meta-materials are an important emerging class of engineered materials in which
complex macroscopic behaviour--whether electromagnetic, thermal, or mechanical--
arises from modular substructure.
Simulation and optimization of these materials are computationally challenging,
as rich substructures necessitate high-fidelity finite element meshes to solve t
he governing PDEs.
To address this, we leverage parametric modular structure to learn component-lev
el surrogates, enabling cheaper high-fidelity simulation.
We use a neural network to model the stored potential energy in a component give
n boundary conditions. This yields a structured prediction task: macroscopic beh
avior is determined by the minimizer of the system's total potential energy, whi
ch can be approximated by composing these surrogate models. Composable energy su
rrogates thus permit simulation in the reduced basis of component boundaries. Co
stly ground-truth simulation of the full structure is avoided, as
training data are generated by performing finite element analysis of individual
components. Using dataset aggregation to choose training data allows us to learn
 energy surrogates which produce accurate macroscopic behavior when composed, ac
celerating simulation of parametric meta-materials.
**********

Efficient Contextual Bandits with Continuous Actions

Maryam Majzoubi, Chicheng Zhang, Rajan Chari, Akshay Krishnamurthy, John Langfor
d, Aleksandrs Slivkins

We create a computationally tractable learning algorithm for contextual bandits
with continuous actions having unknown structure.  The new reduction-style algor

ithm composes with most supervised learning representations.  We prove that this algorithm works in a general sense and verify the new functionality with large-scale experiments.
**********

Achieving Equalized Odds by Resampling Sensitive Attributes
Yaniv Romano, Stephen Bates, Emmanuel Candes
We present a flexible framework for learning predictive models that approximately satisfy the equalized odds notion of fairness. This is achieved by introducing a general discrepancy functional that rigorously quantifies violations of this criterion. This differentiable functional is used as a penalty driving the model parameters towards equalized odds. To rigorously evaluate fitted models, we develop a formal hypothesis test to detect whether a prediction rule violates this property, the first such test in the literature. Both the model fitting and hypothesis testing leverage a resampled version of the sensitive attribute obeying equalized odds, by construction. We demonstrate the applicability and validity of the proposed framework both in regression and multi-class classification problems, reporting improved performance over state-of-the-art methods. Lastly, we show how to incorporate techniques for equitable uncertainty quantification---unbiased for each group under study---to communicate the results of the data analysis in exact terms.
**********

Multi-Robot Collision Avoidance under Uncertainty with Probabilistic Safety Barrier Certificates
Wenhao Luo, Wen Sun, Ashish Kapoor
Safety in terms of collision avoidance for multi-robot systems is a difficult challenge under uncertainty, non-determinism, and lack of complete information. This paper aims to propose a collision avoidance method that accounts for both measurement uncertainty and motion uncertainty. In particular, we propose Probabilistic Safety Barrier Certificates (PrSBC) using Control Barrier Functions to define the space of admissible control actions that are probabilistically safe with formally provable theoretical guarantee. By formulating the chance constrained safety set into deterministic control constraints with PrSBC, the method entails minimally modifying an existing controller to determine an alternative safe controller via quadratic programming constrained to PrSBC constraints. The key advantage of the approach is that no assumptions about the form of uncertainty are required other than finite support, also enabling worst-case guarantees. We demonstrate effectiveness of the approach through experiments on realistic simulation environments.
**********

Hard Shape-Constrained Kernel Machines
Pierre-Cyril Aubin-Frankowski, Zoltan Szabo
Shape constraints (such as non-negativity, monotonicity, convexity) play a central role in a large number of applications, as they usually improve performance for small sample size and help interpretability. However enforcing these shape requirements in a hard fashion is an extremely challenging problem. Classically, this task is tackled (i) in a soft way (without out-of-sample guarantees), (ii) by specialized transformation of the variables on a case-by-case basis, or (iii) by using highly restricted function classes, such as polynomials or polynomial splines. In this paper, we prove that hard affine shape constraints on function derivatives can be encoded in kernel machines which represent one of the most flexible and powerful tools in machine learning and statistics. Particularly, we present a tightened second-order cone constrained reformulation, that can be readily implemented in convex solvers. We prove performance guarantees on the solution, and demonstrate the efficiency of the approach in joint quantile regression with applications to economics and to the analysis of aircraft trajectories, among others.
**********

A Closer Look at the Training Strategy for Modern Meta-Learning
JIAXIN CHEN, Xiao-Ming Wu, Yanke Li, Qimai LI, Li-Ming Zhan, Fu-lai Chung
Requests for name changes in the electronic proceedings will be accepted with no

questions asked.  However name changes may cause bibliographic tracking issues.  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 Value of Out-of-Distribution Testing: An Example of Goodhart's Law
Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, Anton van den Hengel
Out-of-distribution (OOD) testing is increasingly popular for evaluating a machine learning system's ability to generalize beyond the biases of a training set. OOD benchmarks are designed to present a different joint distribution of data and labels between training and test time. VQA-CP has become the standard OOD benchmark for visual question answering, but we discovered three troubling practices in its current use. First, most published methods rely on explicit knowledge of the construction of the OOD splits. They often rely on inverting'' the distribution of labels, e.g. answering mostlyyes'' when the common training answer was ``no''. Second, the OOD test set is used for model selection. Third, a model's in-domain performance is assessed after retraining it on in-domain splits (VQA v2) that exhibit a more balanced distribution of labels. These three practices defeat the objective of evaluating generalization, and put into question the value of methods specifically designed for this dataset. We show that embarrassingly-simple methods, including one that generates answers at random, surpass the state of the art on some question types. We provide short- and long-term solutions to avoid these pitfalls and realize the benefits of OOD evaluation.
**********

Generalised Bayesian Filtering via Sequential Monte Carlo
Ayman Boustati, Omer Deniz Akyildiz, Theodoros Damoulas, Adam Johansen
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Deterministic Approximation for Submodular Maximization over a Matroid in Nearly Linear Time
Kai Han, zongmai Cao, Shuang Cui, Benwei 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.
**********

Flows for simultaneous manifold learning and density estimation
Johann Brehmer, Kyle Cranmer
We introduce manifold-learning flows (■-flows), a new class of generative models that simultaneously learn the data manifold as well as a tractable probability density on that manifold. Combining aspects of normalizing flows, GANs, autoencoders, and energy-based models, they have the potential to represent data sets with a manifold structure more faithfully and provide handles on dimensionality reduction, denoising, and out-of-distribution detection. We argue why such models should not be trained by maximum likelihood alone and present a new training algorithm that separates manifold and density updates. In a range of experiments we demonstrate how ■-flows learn the data manifold and allow for better inference than standard flows in the ambient data space.
**********

Simultaneous Preference and Metric Learning from Paired Comparisons
Austin Xu, Mark Davenport
A popular model of preference in the context of recommendation systems is the so-called ideal point model. In this model, a user is represented as a vector u together with a collection of items x1 ... xN in a common low-dimensional space. The vector u represents the user's "ideal point," or the ideal combination of features that represents a hypothesized most preferred item. The underlying assumption in this model is that a smaller distance between u and an item xj  indicates

a stronger preference for xj. In the vast majority of the existing work on learning ideal point models, the underlying distance has been assumed to be Euclidean. However, this eliminates any possibility of interactions between features and a user's underlying preferences. In this paper, we consider the problem of learning an ideal point representation of a user's preferences when the distance metric is an unknown Mahalanobis metric. Specifically, we present a novel approach to estimate the user's ideal point u and the Mahalanobis metric from paired comparisons of the form "item xi is preferred to item xj.'' This can be viewed as a special case of a more general metric learning problem where the location of some points are unknown a priori. We conduct extensive experiments on synthetic and real-world datasets to exhibit the effectiveness of our algorithm.
**********

Efficient Variational Inference for Sparse Deep Learning with Theoretical Guarantee

Jincheng Bai, Qifan Song, Guang Cheng

Sparse deep learning aims to address the challenge of huge storage consumption by deep neural networks, and to recover the sparse structure of target functions. Although tremendous empirical successes have been achieved, most sparse deep learning algorithms are lacking of theoretical supports. On the other hand, another line of works have proposed theoretical frameworks that are computationally infeasible. In this paper, we train sparse deep neural networks with a fully Bayesian treatment under spike-and-slab priors, and develop a set of computationally efficient variational inferences via continuous relaxation of Bernoulli distribution. The variational posterior contraction rate is provided, which justifies the consistency of the proposed variational Bayes method. Interestingly, our empirical results demonstrate that this variational procedure provides uncertainty quantification in terms of Bayesian predictive distribution and is also capable to accomplish consistent variable selection by training a sparse multi-layer neural network.
**********

Learning Manifold Implicitly via Explicit Heat-Kernel Learning

Yufan Zhou, Changyou Chen, Jinhui Xu

Manifold learning is a fundamental problem in machine learning with numerous applications. Most of the existing methods directly learn the low-dimensional embedding of the data in some high-dimensional space, and usually lack the flexibility of being directly applicable to down-stream applications. In this paper, we propose the concept of implicit manifold learning, where manifold information is implicitly obtained by learning the associated heat kernel. A heat kernel is the solution of the corresponding heat equation, which describes how ``heat'' transfers on the manifold, thus containing ample geometric information of the manifold. We provide both practical algorithm and theoretical analysis of our framework. The learned heat kernel can be applied to various kernel-based machine learning models, including deep generative models (DGM) for data generation and Stein Variational Gradient Descent for Bayesian inference. Extensive experiments show that our framework can achieve the state-of-the-art results compared to existing methods for the two tasks.
**********

Deep Relational Topic Modeling via Graph Poisson Gamma Belief Network

Chaojie Wang, Hao Zhang, Bo Chen, Dongsheng Wang, Zhengjue Wang, Mingyuan Zhou

To analyze a collection of interconnected documents, relational topic models (RTMs) have been developed to describe both the link structure and document content, exploring their underlying relationships via a single-layer latent representation with limited expressive capability. To better utilize the document network, we first propose graph Poisson factor analysis (GPFA) that constructs a probabilistic model for interconnected documents and also provides closed-form Gibbs sampling update equations, moving beyond sophisticated approximate assumptions of existing RTMs. Extending GPFA, we develop a novel hierarchical RTM named graph Poisson gamma belief network (GPGBN), and further introduce two different Weibull distribution based variational graph auto-encoders for efficient model inference and effective network information aggregation. Experimental results demonstrate

that our models extract high-quality hierarchical latent document representatio
ns, leading to improved performance over baselines on various graph analytic tas
ks.
**********

One-bit Supervision for Image Classification
Hengtong Hu, Lingxi Xie, Zewei Du, Richang Hong, Qi Tian
This paper presents one-bit supervision, a novel setting of learning from incomp
lete annotations, in the scenario of image classification. Instead of training a
 model upon the accurate label of each sample, our setting requires the model to
 query with a predicted label of each sample and learn from the answer whether t
he guess is correct. This provides one bit (yes or no) of information, and more
importantly, annotating each sample becomes much easier than finding the accurat
e label from many candidate classes. There are two keys to training a model upon
 one-bit supervision: improving the guess accuracy and making use of incorrect g
uesses. For these purposes, we propose a multi-stage training paradigm which inc
orporates negative label suppression into an off-the-shelf semi-supervised learn
ing algorithm. In three popular image classification benchmarks, our approach cl
aims higher efficiency in utilizing the limited amount of annotations.
**********

What is being transferred in transfer learning?
Behnam Neyshabur, Hanie Sedghi, Chiyuan Zhang
One desired capability for machines is the ability to transfer their understandi
ng of one domain to another domain where data is (usually) scarce. Despite ample
 adaptation of transfer learning in many deep learning applications, we yet do n
ot understand what enables a successful transfer and which part of the network i
s responsible for that. In this paper, we provide new tools and analysis to addr
ess these fundamental questions. Through a series of analysis on transferring to
 block-shuffled images, we separate the effect of feature reuse from learning hi
gh-level statistics of data and show that some benefit of transfer learning come
s from the latter. We present that when training from pre-trained weights, the m
odel stays in the same basin in the loss landscape and different instances of su
ch model are similar in feature space and close in parameter space.
**********

Submodular Maximization Through Barrier Functions
Ashwinkumar Badanidiyuru, Amin Karbasi, Ehsan Kazemi, Jan Vondrak
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Neural Networks with Recurrent Generative Feedback
Yujia Huang, James Gornet, Sihui Dai, Zhiding Yu, Tan Nguyen, Doris Tsao, Anima
Anandkumar
Neural networks are vulnerable to input perturbations such as additive noise and
 adversarial attacks. In contrast, human perception is much more robust to such
perturbations. The Bayesian brain hypothesis states that human brains use an int
ernal generative model to update the posterior beliefs of the sensory input. Thi
s mechanism can be interpreted as a form of self-consistency between the maximum
 a posteriori (MAP) estimation of an internal generative model and the external
environment. Inspired by such hypothesis, we enforce self-consistency in neural
networks by incorporating generative recurrent feedback. We instantiate this des
ign on convolutional neural networks (CNNs). The proposed framework, termed Conv
olutional Neural Networks with Feedback (CNN-F), introduces a generative feedbac
k with latent variables to existing CNN architectures, where consistent predicti
ons are made through alternating MAP inference under a Bayesian framework. In th
e experiments, CNN-F shows considerably improved adversarial robustness over con
ventional feedforward CNNs on standard benchmarks.
**********

Learning to Extrapolate Knowledge: Transductive Few-shot Out-of-Graph Link Predi
ction

Jinheon Baek, Dong Bok Lee, Sung Ju Hwang
Many practical graph problems, such as knowledge graph construction and drug-dru g interaction prediction, require to handle multi-relational graphs. However, ha ndling real-world multi-relational graphs with Graph Neural Networks (GNNs) is o ften challenging due to their evolving nature, as new entities (nodes) can emerg e over time. Moreover, newly emerged entities often have few links, which makes the learning even more difficult. Motivated by this challenge, we introduce a re alistic problem of few-shot out-of-graph link prediction, where we not only pred ict the links between the seen and unseen nodes as in a conventional out-of-know ledge link prediction task but also between the unseen nodes, with only few edge s per node. We tackle this problem with a novel transductive meta-learning frame work which we refer to as Graph Extrapolation Networks (GEN). GEN meta-learns bo th the node embedding network for inductive inference (seen-to-unseen) and the l ink prediction network for transductive inference (unseen-to-unseen). For transd uctive link prediction, we further propose a stochastic embedding layer to model uncertainty in the link prediction between unseen entities. We validate our mod el on multiple benchmark datasets for knowledge graph completion and drug-drug i nteraction prediction. The results show that our model significantly outperforms relevant baselines for out-of-graph link prediction tasks.
**********

Exploiting weakly supervised visual patterns to learn from partial annotations
Kaustav Kundu, Joseph Tighe
As classifications datasets progressively get larger in terms of label space and number of examples, annotating them with all labels becomes non-trivial and exp ensive task. For example, annotating the entire OpenImage test set can cost $6.5 M. Hence, in current large-scale benchmarks such as OpenImages and LVIS, less th an 1\% of the labels are annotated across all images. Standard classification mo dels are trained in a manner where these un-annotated labels are ignored. Ignori ng these un-annotated labels result in loss of supervisory signal which reduces the performance of the classification models. Instead, in this paper, we exploit relationships among images and labels to derive more supervisory signal from th e un-annotated labels.  We study the effectiveness of our approach across severa l multi-label computer vision benchmarks, such as CIFAR100, MS-COCO panoptic seg mentation, OpenImage and LVIS datasets. Our approach can outperform baselines by a margin of 2-10% across all the datasets on mean average precision (mAP) and m ean F1 metrics.
**********

Improving Inference for Neural Image Compression
Yibo Yang, Robert Bamler, Stephan Mandt
We consider the problem of lossy image compression with deep latent variable mod els. State-of-the-art methods build on hierarchical variational autoencoders (VA Es) and learn inference networks to predict a compressible latent representation of each data point. Drawing on the variational inference perspective on compres sion, we identify three approximation gaps which limit performance in the conven tional approach: an amortization gap, a discretization gap, and a marginalizatio n gap. We propose remedies for each of these three limitations based on ideas re lated to iterative inference, stochastic annealing for discrete optimization, an d bits-back coding, resulting in the first application of bits-back coding to lo ssy compression. In our experiments, which include extensive baseline comparison s and ablation studies, we achieve new state-of-the-art performance on lossy ima ge compression using an established VAE architecture, by changing only the infer ence method.
**********

Neuron Merging: Compensating for Pruned Neurons
Woojeong Kim, Suhyun Kim, Mincheol Park, Geunseok Jeon
Network pruning is widely used to lighten and accelerate neural network models. Structured network pruning discards the whole neuron or filter, leading to accur acy loss. In this work, we propose a novel concept of neuron merging applicable to both fully connected layers and convolution layers, which compensates for the information loss due to the pruned neurons/filters. Neuron merging starts with

decomposing the original weights into two matrices/tensors. One of them becomes the new weights for the current layer, and the other is what we name a scaling matrix, guiding the combination of neurons. If the activation function is ReLU, the scaling matrix can be absorbed into the next layer under certain conditions, compensating for the removed neurons. We also propose a data-free and inexpensive method to decompose the weights by utilizing the cosine similarity between neurons. Compared to the pruned model with the same topology, our merged model better preserves the output feature map of the original model; thus, it maintains the accuracy after pruning without fine-tuning. We demonstrate the effectiveness of our approach over network pruning for various model architectures and datasets. As an example, for VGG-16 on CIFAR-10, we achieve an accuracy of 93.16% while reducing 64% of total parameters, without any fine-tuning. The code can be found here: https://github.com/friendshipkim/neuron-merging
**********

FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence
Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, Chun-Liang Li
Semi-supervised learning (SSL) provides an effective means of leveraging unlabeled data to improve a model’s performance. This domain has seen fast progress recently, at the cost of requiring more complex methods. In this paper we propose FixMatch, an algorithm that is a significant simplification of existing SSL methods. FixMatch first generates pseudo-labels using the model’s predictions on weakly-augmented unlabeled images. For a given image, the pseudo-label is only retained if the model produces a high-confidence prediction. The model is then trained to predict the pseudo-label when fed a strongly-augmented version of the same image. Despite its simplicity, we show that FixMatch achieves state-of-the-art performance across a variety of standard semi-supervised learning benchmarks, including 94.93% accuracy on CIFAR-10 with 250 labels and 88.61% accuracy with 40 – just 4 labels per class. We carry out an extensive ablation study to tease apart the experimental factors that are most important to FixMatch’s success. The code is available at https://github.com/google-research/fixmatch.
**********

Reinforcement Learning with Combinatorial Actions: An Application to Vehicle Routing
Arthur Delarue, Ross Anderson, Christian Tjandraatmadja
Value-function-based methods have long played an important role in reinforcement learning. However, finding the best next action given a value function of arbitrary complexity is nontrivial when the action space is too large for enumeration. We develop a framework for value-function-based deep reinforcement learning with a combinatorial action space, in which the action selection problem is explicitly formulated as a mixed-integer optimization problem. As a motivating example, we present an application of this framework to the capacitated vehicle routing problem (CVRP), a combinatorial optimization problem in which a set of locations must be covered by a single vehicle with limited capacity. On each instance, we model an action as the construction of a single route, and consider a deterministic policy which is improved through a simple policy iteration algorithm. Our approach is competitive with other reinforcement learning methods and achieves an average gap of 1.7% with state-of-the-art OR methods on standard library instances of medium size.
**********

Towards Playing Full MOBA Games with Deep Reinforcement Learning
Deheng Ye, Guibin Chen, Wen Zhang, Sheng Chen, Bo Yuan, Bo Liu, Jia Chen, Zhao Liu, Fuhao Qiu, Hongsheng Yu, Yinyuting Yin, Bei Shi, Liang Wang, Tengfei Shi, Qiang Fu, Wei Yang, Lanxiao Huang, Wei Liu
MOBA games, e.g., Honor of Kings, League of Legends, and Dota 2, pose grand challenges to AI systems such as multi-agent, enormous state-action space, complex action control, etc. Developing AI for playing MOBA games has raised much attention accordingly. However, existing work falls short in handling the raw game complexity caused by the explosion of agent combinations, i.e., lineups, when expanding the hero pool in case that OpenAI's Dota AI limits the play to a pool of onl

y 17 heroes. As a result, full MOBA games without restrictions are far from bein
g mastered by any existing AI system. In this paper, we propose a MOBA AI learni
ng paradigm that methodologically enables playing full MOBA games with deep rein
forcement learning. Specifically, we develop a combination of novel and existing
 learning techniques, including off-policy adaption, multi-head value estimation
, curriculum self-play learning, policy distillation, and Monte-Carlo tree-searc
h, in training and playing a large pool of heroes, meanwhile addressing the scal
ability issue skillfully. Tested on Honor of Kings, a popular MOBA game, we show
 how to build superhuman AI agents that can defeat top esports players. The supe
riority of our AI is demonstrated by the first large-scale performance test of M
OBA AI agent in the literature.
**********

Rankmax: An Adaptive Projection Alternative to the Softmax Function
Weiwei Kong, Walid Krichene, Nicolas Mayoraz, Steffen Rendle, Li Zhang
Several machine learning models involve mapping a score vector to a probability
vector. Usually, this is done by projecting the score vector onto a probability
simplex, and such projections are often characterized as Lipschitz continuous ap
proximations of the argmax function, whose Lipschitz constant is controlled by a
 parameter that is similar to a softmax temperature. The aforementioned paramete
r has been observed to affect the quality of these models and is typically eithe
r treated as a constant or decayed over time. In this work, we propose a method
that adapts this parameter to individual training examples. The resulting method
 exhibits desirable properties, such as sparsity of its support and numerically
efficient implementation, and we find that it significantly outperforms competin
g non-adaptive projection methods. In our analysis, we also derive the general s
olution of (Bregman) projections onto the (n, k)-simplex, a result which may be
of independent interest.
**********

Online Agnostic Boosting via Regret Minimization
Nataly Brukhim, Xinyi Chen, Elad Hazan, Shay Moran
Boosting is a widely used machine learning approach based on the idea of aggrega
ting weak learning
rules. While in statistical learning numerous boosting methods exist both in the
 realizable and agnostic
settings, in online learning they exist only in the realizable case. In this wor
k we provide the first agnostic
online boosting algorithm; that is, given a weak learner with only marginally-be
tter-than-trivial regret
guarantees, our algorithm boosts it to a strong learner with sublinear regret.
Our algorithm is based on an abstract (and simple) reduction to online convex op
timization, which
efficiently converts an arbitrary online convex optimizer to an online booster.
Moreover, this reduction
extends to the statistical as well as the online realizable settings, thus unify
ing the 4 cases of statistical/online and agnostic/realizable boosting.
**********

Causal Intervention for Weakly-Supervised Semantic Segmentation
Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, Qianru Sun
We present a causal inference framework to improve Weakly-Supervised Semantic Se
gmentation (WSSS). Specifically, we aim to generate better pixel-level pseudo-ma
sks by using only image-level labels -- the most crucial step in WSSS. We attrib
ute the cause of the ambiguous boundaries of pseudo-masks to the confounding con
text, e.g., the correct image-level classification of "horse" and "person" may
be not only due to the recognition of each instance, but also their co-occurrenc
e context, making the model inspection (e.g., CAM) hard to distinguish between t
he boundaries. Inspired by this, we propose a structural causal model to analyze
 the causalities among images, contexts, and class labels. Based on it, we devel
op a new method: Context Adjustment (CONTA), to remove the confounding bias in i
mage-level classification and thus provide better pseudo-masks as ground-truth f
or the subsequent segmentation model. On PASCAL VOC 2012 and MS-COCO, we show th

at CONTA boosts various popular WSSS methods to new state-of-the-arts.
**********
Belief Propagation Neural Networks
Jonathan Kuck, Shuvam Chakraborty, Hao Tang, Rachel Luo, Jiaming Song, Ashish Sabharwal, Stefano Ermon
Learned neural solvers have successfully been used to solve combinatorial optimization and decision problems. More general counting variants of these problems, however, are still largely solved with hand-crafted solvers. To bridge this gap, we introduce belief propagation neural networks (BPNNs), a class of parameterized operators that operate on factor graphs and generalize Belief Propagation (BP). In its strictest form, a BPNN layer (BPNN-D) is a learned iterative operator that provably maintains many of the desirable properties of BP for any choice of the parameters. Empirically, we show that by training BPNN-D learns to perform the task better than the original BP: it converges 1.7x faster on Ising models while providing tighter bounds. On challenging model counting problems, BPNNs compute estimates 100's of times faster than state-of-the-art handcrafted methods, while returning an estimate of comparable quality.
**********
Over-parameterized Adversarial Training: An Analysis Overcoming the Curse of Dimensionality
Yi Zhang, Orestis Plevrakis, Simon S. Du, Xingguo Li, Zhao Song, Sanjeev 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.
**********
Post-training Iterative Hierarchical Data Augmentation for Deep Networks
Adil Khan, Khadija Fraz
In this paper, we propose a new iterative hierarchical data augmentation (IHDA) method to fine-tune trained deep neural networks to improve their generalization performance. The IHDA is motivated by three key insights: (1) Deep networks (DNs) are good at learning multi-level representations from data. (2) Performing data augmentation (DA) in the learned feature spaces of DNs can significantly improve their performance. (3) Implementing DA in hard-to-learn regions of a feature space can effectively augment the dataset to improve generalization.  Accordingly, the IHDA performs DA in a deep feature space, at level l, by transforming it into a distribution space and synthesizing new samples using the learned distributions for data points that lie in hard-to-classify regions, which is estimated by analyzing the neighborhood characteristics of each data point. The synthesized samples are used to fine-tune the parameters of the subsequent layers. The same procedure is then repeated for the feature space at level l+1. To avoid overfitting, the concept of dropout probability is employed, which is gradually relaxed as the IHDA works towards high-level feature spaces. IHDA provided a state-of-the-art performance on CIFAR-10, CIFAR-100, and ImageNet for several DNs, and beat the performance of existing state-of-the-art DA approaches for the same networks on these datasets. Finally, to demonstrate its domain-agnostic properties, we show the significant improvements that IHDA provided for a deep neural network on a non-image wearable sensor-based activity recognition benchmark.
**********
Debugging Tests for Model Explanations
Julius Adebayo, Michael Muelly, Ilaria Liccardi, Been Kim
We investigate whether post-hoc model explanations are effective for diagnosing model errors--model debugging. In response to the challenge of explaining a model's prediction, a vast array of explanation methods have been proposed.  Despite increasing use, it is unclear if they are effective. To start, we categorize \textit{bugs}, based on their source, into: ~\textit{data, model, and test-time} contamination bugs. For several explanation methods, we assess their ability to: detect spurious correlation artifacts (data contamination), diagnose mislabeled training examples (data contamination), differentiate between a (partially) re-initialized model and a trained one (model contamination), and detect out-of-dist

ribution inputs (test-time contamination). We find that the methods tested are a ble to diagnose a spurious background bug, but not conclusively identify mislabe led training examples. In addition, a class of methods, that modify the back-pro pagation algorithm are invariant to the higher layer parameters of a deep networ k; hence, ineffective for diagnosing model contamination. We complement our anal ysis with a human subject study, and find that subjects fail to identify defecti ve models using attributions, but instead rely, primarily, on model predictions. Taken together, our results provide guidance for practitioners and researchers turning to explanations as tools for model debugging.
**********

Robust compressed sensing using generative models
Ajil Jalal, Liu Liu, Alexandros G. Dimakis, Constantine Caramanis
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Fairness without Demographics through Adversarially Reweighted Learning
Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, Ed Chi
Much of the previous machine learning (ML) fairness literature assumes that prot ected features such as race and sex are present in the dataset, and relies upon them to mitigate fairness concerns. However, in practice factors like privacy an d regulation often preclude the collection of protected features, or their use f or training or inference, severely limiting the applicability of traditional fai rness research. Therefore, we ask: How can we train a ML model to improve fairne ss when we do not even know the protected group memberships? In this work we add ress this problem by proposing Adversarially Reweighted Learning (ARL). In parti cular, we hypothesize that non-protected features and task labels are valuable f or identifying fairness issues, and can be used to co-train an adversarial rewei ghting approach for improving fairness. Our results show that ARL improves Rawls ian Max-Min fairness, with notable AUC improvements for worst-case protected gro ups in multiple datasets, outperforming state-of-the-art alternatives.
**********

Stochastic Latent Actor-Critic: Deep Reinforcement Learning with a Latent Variab le Model
Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, Sergey Levine
Deep reinforcement learning (RL) algorithms can use high-capacity deep networks to learn directly from image observations. However, these high-dimensional obser vation spaces present a number of challenges in practice, since the policy must now solve two problems: representation learning and task learning. In this work , we tackle these two problems separately, by explicitly learning latent represe ntations that can accelerate reinforcement learning from images. We propose the stochastic latent actor-critic (SLAC) algorithm: a sample-efficient and high-per forming RL algorithm for learning policies for complex continuous control tasks directly from high-dimensional image inputs. SLAC provides a novel and principle d approach for unifying stochastic sequential models and RL into a single method , by learning a compact latent representation and then performing RL in the mode l's learned latent space. Our experimental evaluation demonstrates that our meth od outperforms both model-free and model-based alternatives in terms of final pe rformance and sample efficiency, on a range of difficult image-based control tas ks. Our code and videos of our results are available at our website.
**********

Ridge Rider: Finding Diverse Solutions by Following Eigenvectors of the Hessian
Jack Parker-Holder, Luke Metz, Cinjon Resnick, Hengyuan Hu, Adam Lerer, Alistair Letcher, Alexander Peysakhovich, Aldo Pacchiano, Jakob Foerster
Over the last decade, a single algorithm has changed many facets of our lives - Stochastic Gradient Descent (SGD). In the era of ever decreasing loss functions, SGD and its various offspring have become the go-to optimization tool in machin e learning and are a key component of the success of deep neural networks (DNNs)

. While SGD is guaranteed to converge to a local optimum (under loose assumption
s), in some cases it may matter which local optimum is found, and this is often
context-dependent. Examples frequently arise in machine learning, from shape-ver
sus-texture-features to ensemble methods and zero-shot coordination.  In these s
ettings, there are desired solutions which SGD on standard' loss functions will
not find, since it instead converges to theeasy' solutions.  In this paper, we p
resent a different approach. Rather than following the gradient, which correspon
ds to a locally greedy direction, we instead follow the eigenvectors of the Hess
ian. By iteratively following and branching amongst the ridges, we effectively s
pan the loss surface to find qualitatively different solutions. We show both the
oretically and experimentally that our method, called Ridge Rider (RR), offers a
 promising direction for a variety of challenging problems.
**********

The route to chaos in routing games: When is price of anarchy too optimistic?
Thiparat Chotibut, Fryderyk Falniowski, Micha■ Misiurewicz, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Online Algorithm for Unsupervised Sequential Selection with Contextual Informati
on
Arun Verma, Manjesh Kumar Hanawal, Csaba Szepesvari, Venkatesh Saligrama
In this paper, we study Contextual Unsupervised Sequential Selection (USS), a ne
w variant of the stochastic contextual bandits problem where the loss of an arm
cannot be inferred from the observed feedback. In our setup, arms are associated
 with fixed costs and are ordered, forming a cascade. In each round, a context i
s presented, and the learner selects the arms sequentially till some depth. The
total cost incurred by stopping at an arm is the sum of fixed costs of arms sele
cted and the stochastic loss associated with the arm. The learner's goal is to l
earn a decision rule that maps contexts to arms with the goal of minimizing the
total expected loss. The problem is challenging as we are faced with an unsuperv
ised setting as the total loss cannot be estimated. Clearly, learning is feasibl
e only if the optimal arm can be inferred (explicitly or implicitly) from the pr
oblem structure. We observe that learning is still possible when the problem ins
tance satisfies the so-called 'Contextual Weak Dominance' (CWD) property. Under
CWD, we propose an algorithm for the contextual USS problem and demonstrate that
 it has sub-linear regret. Experiments on synthetic and real datasets validate o
ur algorithm.
**********

Adapting Neural Architectures Between Domains
Yanxi Li, Zhaohui Yang, Yunhe Wang, Chang Xu
Neural architecture search (NAS) has demonstrated impressive performance in auto
matically designing high-performance neural networks. The power of deep neural n
etworks is to be unleashed for analyzing a large volume of data (e.g. ImageNet),
 but the architecture search is often executed on another smaller dataset (e.g.
CIFAR-10) to finish it in a feasible time. However, it is hard to guarantee that
 the optimal architecture derived on the proxy task could maintain its advantage
s on another more challenging dataset. This paper aims to improve the generaliza
tion of neural architectures via domain adaptation. We analyze the generalizatio
n bounds of the derived architecture and suggest its close relations with the va
lidation error and the data distribution distance on both domains. These theoret
ical analyses lead to AdaptNAS, a novel and principled approach to adapt neural
architectures between domains in NAS. Our experimental evaluation shows that onl
y a small part of ImageNet will be sufficient for AdaptNAS to extend its archite
cture success to the entire ImageNet and outperform state-of-the-art comparison
algorithms.
**********

What went wrong and when? Instance-wise feature importance for time-series black
-box models

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K. Duvenaud, Anna Goldenberg

Explanations of time series models are useful for high stakes applications like healthcare but have received little attention in machine learning literature. We propose FIT, a framework that evaluates the importance of observations for a multivariate time-series black-box model by quantifying the shift in the predictive distribution over time. FIT defines the importance of an observation based on its contribution to the distributional shift under a KL-divergence that contrasts the predictive distribution against a counterfactual where the rest of the features are unobserved. We also demonstrate the need to control for time-dependent distribution shifts. We compare with state-of-the-art baselines on simulated and real-world clinical data and demonstrate that our approach is superior in identifying important time points and observations throughout the time series.

**********

Towards Better Generalization of Adaptive Gradient Methods

Yingxue Zhou, Belhal Karimi, Jinxing Yu, Zhiqiang Xu, Ping Li

Adaptive gradient methods such as AdaGrad, RMSprop and Adam have been optimizers of choice for deep learning due to their fast training speed. However, it was recently observed that their generalization performance is often worse than that of SGD for over-parameterized neural networks. While new algorithms such as AdaBound, SWAT, and Padam were proposed to improve the situation, the provided analyses are only committed to optimization bounds for the training objective, leaving critical generalization capacity unexplored. To close this gap, we propose \textit{\textbf{S}table \textbf{A}daptive \textbf{G}radient \textbf{D}escent} (\textsc{SAGD}) for nonconvex optimization which leverages differential privacy to boost the generalization performance of adaptive gradient methods. Theoretical analyses show that \textsc{SAGD} has high-probability convergence to a population stationary point. We further conduct experiments on various popular deep learning tasks and models. Experimental results illustrate that \textsc{SAGD} is empirically competitive and often better than baselines.

**********

Learning Guidance Rewards with Trajectory-space Smoothing

Tanmay Gangwani, Yuan Zhou, Jian Peng

Long-term temporal credit assignment is an important challenge in deep reinforcement learning (RL). It refers to the ability of the agent to attribute actions to consequences that may occur after a long time interval. Existing policy-gradient and Q-learning algorithms typically rely on dense environmental rewards that provide rich short-term supervision and help with credit assignment. However, they struggle to solve tasks with delays between an action and the corresponding rewarding feedback. To make credit assignment easier, recent works have proposed algorithms to learn dense "guidance" rewards that could be used in place of the sparse or delayed environmental rewards. This paper is in the same vein -- starting with a surrogate RL objective that involves smoothing in the trajectory-space, we arrive at a new algorithm for learning guidance rewards. We show that the guidance rewards have an intuitive interpretation, and can be obtained without training any additional neural networks. Due to the ease of integration, we use the guidance rewards in a few popular algorithms (Q-learning, Actor-Critic, Distributional-RL) and present results in single-agent and multi-agent tasks that elucidate the benefit of our approach when the environmental rewards are sparse or delayed.

**********

Variance Reduction via Accelerated Dual Averaging for Finite-Sum Optimization

Chaobing Song, Yong Jiang, Yi 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.

**********

Tree! I am no Tree! I am a low dimensional Hyperbolic Embedding

Rishi Sonthalia, Anna Gilbert

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Structural Causal Models for Tractable Counterfactual Inference
Nick Pawlowski, Daniel Coelho de Castro, Ben Glocker
We formulate a general framework for building structural causal models (SCMs) with deep learning components. The proposed approach employs normalising flows and variational inference to enable tractable inference of exogenous noise variables - a crucial step for counterfactual inference that is missing from existing deep causal learning methods. Our framework is validated on a synthetic dataset built on MNIST as well as on a real-world medical dataset of brain MRI scans. Our experimental results indicate that we can successfully train deep SCMs that are capable of all three levels of Pearl's ladder of causation: association, intervention, and counterfactuals, giving rise to a powerful new approach for answering causal questions in imaging applications and beyond.
**********

Convolutional Generation of Textured 3D Meshes
Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, Aurelien Lucchi
While recent generative models for 2D images achieve impressive visual results, they clearly lack the ability to perform 3D reasoning. This heavily restricts the degree of control over generated objects as well as the possible applications of such models. In this work, we bridge this gap by leveraging recent advances in differentiable rendering. We design a framework that can generate triangle meshes and associated high-resolution texture maps, using only 2D supervision from single-view natural images. A key contribution of our work is the encoding of the mesh and texture as 2D representations, which are semantically aligned and can be easily modeled by a 2D convolutional GAN. We demonstrate the efficacy of our method on Pascal3D+ Cars and CUB, both in an unconditional setting and in settings where the model is conditioned on class labels, attributes, and text. Finally, we propose an evaluation methodology that assesses the mesh and texture quality separately.
**********

A Statistical Framework for Low-bitwidth Training of Deep Neural Networks
Jianfei Chen, Yu Gai, Zhewei Yao, Michael W. Mahoney, Joseph E. Gonzalez
Fully quantized training (FQT), which uses low-bitwidth hardware by quantizing the activations, weights, and gradients of a neural network model, is a promising approach to accelerate the training of deep neural networks. One major challenge with FQT is the lack of theoretical understanding, in particular of how gradient quantization impacts convergence properties. In this paper, we address this problem by presenting a statistical framework for analyzing FQT algorithms. We view the quantized gradient of FQT as a stochastic estimator of its full precision counterpart, a procedure known as quantization-aware training (QAT). We show that the FQT gradient is an unbiased estimator of the QAT gradient, and we discuss the impact of gradient quantization on its variance. Inspired by these theoretical results, we develop two novel gradient quantizers, and we show that these have smaller variance than the existing per-tensor quantizer. For training ResNet-50 on ImageNet, our 5-bit block Householder quantizer achieves only 0.5% validation accuracy loss relative to QAT, comparable to the existing INT8 baseline.
**********

Better Set Representations For Relational Reasoning
Qian Huang, Horace He, Abhay Singh, Yan Zhang, Ser Nam Lim, Austin R. Benson
Incorporating relational reasoning into neural networks has greatly expanded their capabilities and scope. One defining trait of relational reasoning is that it operates on a set of entities, as opposed to standard vector representations. Existing end-to-end approaches for relational reasoning typically extract entities from inputs by directly interpreting the latent feature representations as a set. We show that these approaches do not respect set permutational invariance an

d thus have fundamental representational limitations. To resolve this limitation , we propose a simple and general network module called Set Refiner Network (SRN ). We first use synthetic image experiments to demonstrate how our approach effe ctively decomposes objects without explicit supervision. Then, we insert our mod ule into existing relational reasoning models and show that respecting set invar iance leads to substantial gains in prediction performance and robustness on sev eral relational reasoning tasks. Code can be found at github.com/CUAI/BetterSetR epresentations.
**********

AutoSync: Learning to Synchronize for Data-Parallel Distributed Deep Learning
Hao Zhang, Yuan Li, Zhijie Deng, Xiaodan Liang, Lawrence Carin, Eric Xing
Synchronization is a key step in data-parallel distributed machine learning (ML) . Different synchronization systems and strategies perform differently, and to a chieve optimal parallel training throughput requires synchronization strategies that adapt to model structures and cluster configurations. Existing synchronizat ion systems often only consider a single or a few synchronization aspects, and t he burden of deciding the right synchronization strategy is then placed on the M L practitioners, who may lack the required expertise. In this paper, we develop a model- and resource-dependent representation for synchronization, which unifie s multiple synchronization aspects ranging from architecture, message partitioni ng, placement scheme, to communication topology. Based on this representation, w e build an end-to-end pipeline, AutoSync, to automatically optimize synchronizat ion strategies given model structures and resource specifications, lowering the bar for data-parallel distributed ML. By learning from low-shot data collected i n only 200 trial runs, AutoSync can discover synchronization strategies up to 1. 6x better than manually optimized ones. We develop transfer-learning mechanisms to further reduce the auto-optimization cost --  the simulators can transfer amo ng similar model architectures, among similar cluster configurations, or both. W e also present a dataset that contains over 10000 synchronization strategies and  run-time pairs on a diverse set of models and cluster specifications.
**********

A Combinatorial Perspective on Transfer Learning
Jianan Wang, Eren Sezener, David Budden, Marcus Hutter, Joel Veness
Human intelligence is characterized not only by the capacity to learn complex sk ills, but the ability to rapidly adapt and acquire new skills within an ever-cha nging environment. In this work we study how the learning of modular solutions c an allow for effective generalization to both unseen and potentially differently  distributed data. Our main postulate is that the combination of task segmentati on, modular learning and memory-based ensembling can give rise to generalization  on an exponentially growing number of unseen tasks. We provide a concrete insta ntiation of this idea using a combination of: (1) the Forget-Me-Not Process, for  task segmentation and memory based ensembling; and (2) Gated Linear Networks, w hich in contrast to contemporary deep learning techniques use a modular and loca l learning mechanism. We demonstrate that this system exhibits a number of desir able continual learning properties: robustness to catastrophic forgetting, no ne gative transfer and increasing levels of positive transfer as more tasks are see n. We show competitive performance against both offline and online methods on st andard continual learning benchmarks.
**********

Hardness of Learning Neural Networks with Natural Weights
Amit Daniely, Gal Vardi
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Higher-Order Spectral Clustering of Directed Graphs
Steinar Laenen, He Sun
Clustering is an important topic in algorithms, and has a number of applications  in machine learning, computer vision, statistics, and several other research di

sciplines. Traditional objectives of graph clustering are to find clusters with low conductance. Not only are these objectives just applicable for undirected graphs, they are also incapable to take the relationships between clusters into account, which could be crucial for many applications. To overcome these downsides, we study directed graphs (digraphs) whose clusters exhibit further “structural” information amongst each other. Based on the Hermitian matrix representation of digraphs, we present a nearly-linear time algorithm for digraph clustering, and further show that our proposed algorithm can be implemented in sublinear time under reasonable assumptions. The significance of our theoretical work is demonstrated by extensive experimental results on the UN Comtrade Dataset: the output clustering of our algorithm exhibits not only how the clusters (sets of countries) relate to each other with respect to their import and export records, but also how these clusters evolve over time, in accordance with known facts in international trade.

**********

Primal-Dual Mesh Convolutional Neural Networks

Francesco Milano, Antonio Loquercio, Antoni Rosinol, Davide Scaramuzza, Luca Carlone

Recent works in geometric deep learning have introduced neural networks that allow performing inference tasks on three-dimensional geometric data by defining convolution --and sometimes pooling-- operations on triangle meshes. These methods, however, either consider the input mesh as a graph, and do not exploit specific geometric properties of meshes for feature aggregation and downsampling, or are specialized for meshes, but rely on a rigid definition of convolution that does not properly capture the local topology of the mesh. We propose a method that combines the advantages of both types of approaches, while addressing their limitations: we extend a primal-dual framework drawn from the graph-neural-network literature to triangle meshes, and define convolutions on two types of graphs constructed from an input mesh. Our method takes features for both edges and faces of a 3D mesh as input, and dynamically aggregates them using an attention mechanism. At the same time, we introduce a pooling operation with a precise geometric interpretation, that allows handling variations in the mesh connectivity by clustering mesh faces in a task-driven fashion. We provide theoretical insights of our approach using tools from the mesh-simplification literature. In addition, we validate experimentally our method in the tasks of shape classification and shape segmentation, where we obtain comparable or superior performance to the state of the art.

**********

The Advantage of Conditional Meta-Learning for Biased Regularization and Fine Tuning

Giulia Denevi, Massimiliano Pontil, Carlo Ciliberto

Biased regularization and fine tuning are two recent meta-learning approaches. They have been shown to be effective to tackle distributions of tasks, in which the tasks’ target vectors are all close to a common meta-parameter vector. However, these methods may perform poorly on heterogeneous environments of tasks, where the complexity of the tasks’ distribution cannot be captured by a single meta-parameter vector. We address this limitation by conditional meta-learning, inferring a conditioning function mapping task’s side information into a meta-parameter vector that is appropriate for that task at hand. We characterize properties of the environment under which the conditional approach brings a substantial advantage over standard meta-learning and we highlight examples of environments, such as those with multiple clusters, satisfying these properties. We then propose a convex meta-algorithm providing a comparable advantage also in practice. Numerical experiments confirm our theoretical findings.

**********

Watch out! Motion is Blurring the Vision of Your Deep Neural Networks

Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Bing Yu, Wei Feng, Yang Liu

The state-of-the-art deep neural networks (DNNs) are vulnerable against adversarial examples with additive random-like noise perturbations. While such examples

are hardly found in the physical world, the image blurring effect caused by obje
ct motion, on the other hand, commonly occurs in practice, making the study of w
hich greatly important especially for the widely adopted real-time image process
ing tasks (e.g., object detection, tracking). In this paper, we initiate the fir
st step to comprehensively investigate the potential hazards of blur effect for
DNN, caused by object motion. We propose a novel adversarial attack method that
can generate visually natural motion-blurred adversarial examples, named motion-
based adversarial blur attack (ABBA). To this end, we first formulate the kernel
-prediction-based attack where an input image is convolved with kernels in a pix
el-wise way, and the misclassification capability is achieved by tuning the kern
el weights. To generate visually more natural and plausible examples, we further
 propose the saliency-regularized adversarial kernel prediction, where the salie
nt region serves as a moving object, and the predicted kernel is regularized to
achieve naturally visual effects. Besides, the attack is further enhanced by ada
ptively tuning the translations of object and background. A comprehensive evalua
tion on the NeurIPS'17 adversarial competition dataset demonstrates the effectiv
eness of ABBA by considering various kernel sizes, translations, and regions. Th
e in-depth study further confirms that our method shows a more effective penetra
ting capability to the state-of-the-art GAN-based deblurring mechanisms compared
 with other blurring methods. We release the code to \url{https://github.com/tsi
ngqguo/ABBA}.
**********

Sinkhorn Barycenter via Functional Gradient Descent
Zebang Shen, Zhenfu Wang, Alejandro Ribeiro, Hamed Hassani
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Coresets for Near-Convex Functions
Murad Tukan, Alaa Maalouf, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Bayesian Deep Ensembles via the Neural Tangent Kernel
Bobby He, Balaji Lakshminarayanan, Yee Whye Teh
We explore the link between deep ensembles and Gaussian processes (GPs) through
the lens of the Neural Tangent Kernel (NTK): a recent development in understandi
ng the training dynamics of wide neural networks (NNs). Previous work has shown
that even in the infinite width limit, when NNs become GPs, there is no GP poste
rior interpretation to a deep ensemble trained with squared error loss. We intro
duce a simple modification to standard deep ensembles training, through addition
 of a computationally-tractable, randomised and untrainable function to each ens
emble member, that enables a posterior interpretation in the infinite width limi
t. When ensembled together, our trained NNs give an approximation to a posterior
 predictive distribution, and we prove that our Bayesian deep ensembles make mor
e conservative predictions than standard deep ensembles in the infinite width li
mit. Finally, using finite width NNs we demonstrate that our Bayesian deep ensem
bles faithfully emulate the analytic posterior predictive when available, and ou
tperform standard deep ensembles in various out-of-distribution settings, for bo
th regression and classification tasks.
**********

Improved Schemes for Episodic Memory-based Lifelong Learning
Yunhui Guo, Mingrui Liu, Tianbao Yang, Tajana Rosing
Current deep neural networks can achieve remarkable performance on a single task
. However, when the deep neural network is continually trained on a sequence of
tasks, it seems to gradually forget the previous learned knowledge. This phenome
non is referred to as catastrophic forgetting and motivates the field called lif

elong learning. Recently, episodic memory based approaches such as GEM and A-GEM have shown remarkable performance. In this paper, we provide the first unified view of episodic memory based approaches from an optimization's perspective. This view leads to two improved schemes for episodic memory based lifelong learning, called MEGA-\rom{1} and MEGA-\rom{2}. MEGA-\rom{1} and MEGA-\rom{2} modulate the balance between old tasks and the new task by integrating the current gradient with the gradient computed on the episodic memory. Notably, we show that GEM and A-GEM are degenerate cases of MEGA-\rom{1} and MEGA-\rom{2} which consistently put the same emphasis on the current task, regardless of how the loss changes over time. Our proposed schemes address this issue by using novel loss-balancing updating rules, which drastically improve the performance over GEM and A-GEM. Extensive experimental results show that the proposed schemes significantly advance the state-of-the-art on four commonly used lifelong learning benchmarks, reducing the error by up to 18%.
**********

Adaptive Sampling for Stochastic Risk-Averse Learning
Sebastian Curi, Kfir Y. Levy, Stefanie Jegelka, 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.
**********

Deep Wiener Deconvolution: Wiener Meets Deep Learning for Image Deblurring
Jiangxin Dong, Stefan Roth, Bernt Schiele
We present a simple and effective approach for non-blind image deblurring, combining classical techniques and deep learning. In contrast to existing methods that deblur the image directly in the standard image space, we propose to perform an explicit deconvolution process in a feature space by integrating a classical Wiener deconvolution framework with learned deep features. A multi-scale feature refinement module then predicts the deblurred image from the deconvolved deep features, progressively recovering detail and small-scale structures. The proposed model is trained in an end-to-end manner and evaluated on scenarios with both simulated and real-world image blur. Our extensive experimental results show that the proposed deep Wiener deconvolution network facilitates deblurred results with visibly fewer artifacts. Moreover, our approach quantitatively outperforms state-of-the-art non-blind image deblurring methods by a wide margin.
**********

Discovering Reinforcement Learning Algorithms
Junhyuk Oh, Matteo Hessel, Wojciech M. Czarnecki, Zhongwen Xu, Hado P. van Hasselt, Satinder Singh, David Silver
Reinforcement learning (RL) algorithms update an agent’s parameters according to one of several possible rules, discovered manually through years of research. Automating the discovery of update rules from data could lead to more efficient algorithms, or algorithms that are better adapted to specific environments. Although there have been prior attempts at addressing this significant scientific challenge, it remains an open question whether it is feasible to discover alternatives to fundamental concepts of RL such as value functions and temporal-difference learning. This paper introduces a new meta-learning approach that discovers an entire update rule which includes both what to predict' (e.g. value functions) andhow to learn from it' (e.g. bootstrapping) by interacting with a set of environments. The output of this method is an RL algorithm that we call Learned Policy Gradient (LPG). Empirical results show that our method discovers its own alternative to the concept of value functions. Furthermore it discovers a bootstrapping mechanism to maintain and use its predictions. Surprisingly, when trained solely on toy environments, LPG generalises effectively to complex Atari games and achieves non-trivial performance. This shows the potential to discover general RL algorithms from data.
**********

Taming Discrete Integration via the Boon of Dimensionality
Jeffrey Dudek, Dror Fried, Kuldeep S Meel

Discrete integration is a fundamental problem in computer science that concerns the computation of discrete sums over exponentially large sets. Despite intense interest from researchers for over three decades, the design of scalable techniques for computing estimates with rigorous guarantees for discrete integration remains the holy grail.  The key contribution of this work addresses this scalability challenge via an efficient reduction of discrete integration to model counting. The proposed reduction is achieved via a significant increase in the dimensionality that, contrary to conventional wisdom, leads to solving an instance of the relatively simpler problem of model counting.
**********

Blind Video Temporal Consistency via Deep Video Prior
Chenyang Lei, Yazhou Xing, Qifeng Chen
Applying image processing algorithms independently to each video frame often leads to temporal inconsistency in the resulting video. To address this issue, we present a novel and general approach for blind video temporal consistency. Our method is only trained on a pair of original and processed videos directly instead of a large dataset. Unlike most previous methods that enforce temporal consistency with optical flow, we show that temporal consistency can be achieved by training a convolutional network on a video with the Deep Video Prior. Moreover, a carefully designed iteratively reweighted training strategy is proposed to address the challenging multimodal inconsistency problem. We demonstrate the effectiveness of our approach on 7 computer vision tasks on videos. Extensive quantitative and perceptual experiments show that our approach obtains superior performance than state-of-the-art methods on blind video temporal consistency.
**********

Simplify and Robustify Negative Sampling for Implicit Collaborative Filtering
Jingtao Ding, Yuhan Quan, Quanming Yao, Yong Li, Depeng Jin
Negative sampling approaches are prevalent in implicit collaborative ■ltering for obtaining negative labels from massive unlabeled data. As two major concerns in negative sampling, ef■ciency and effectiveness are still not fully achieved by recent works that use complicate structures and overlook risk of false negative instances. In this paper, we ■rst provide a novel understanding of negative instances by empirically observing that only a few instances are potentially important for model learning, and false negatives tend to have stable predictions over many training iterations. Above ■ndings motivate us to simplify the model by sampling from designed memory that only stores a few important candidates and, more importantly, tackle the untouched false negative problem by favouring high-variance samples stored in memory, which achieves ef■cient sampling of true negatives with high-quality. Empirical results on two synthetic datasets and three real-world datasets demonstrate both robustness and superiorities of our negative sampling method. The implementation is available at https://github.com/dingjingtao/SRNS.
**********

Model Selection for Production System via Automated Online Experiments
Zhenwen Dai, Praveen Chandar, Ghazal Fazelnia, Benjamin Carterette, Mounia Lalmas
A challenge that machine learning practitioners in the industry face is the task of selecting the best model to deploy in production. As a model is often an intermediate component of a production system, online controlled experiments such as A/B tests yield the most reliable estimation of the effectiveness of the whole system, but can only compare two or a few models due to budget constraints. We propose an automated online experimentation mechanism that can efficiently perform model selection from a large pool of models with a small number of online experiments. We derive the probability distribution of the metric of interest that contains the model uncertainty from our Bayesian surrogate model trained using historical logs. Our method efficiently identifies the best model by sequentially selecting and deploying a list of models from the candidate set that balance exploration-exploitation. Using simulations based on real data, we demonstrate the effectiveness of our method on two different tasks.
**********

On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Prob
lems
Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond
Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bha
vya Kailkhura, Xue Lin, Cho-Jui Hsieh
Linear relaxation based perturbation analysis (LiRPA) for neural networks, which
 computes provable linear bounds of output neurons given a certain amount of inp
ut perturbation, has become a core component in robustness verification and cert
ified defense. The majority of LiRPA-based methods focus on simple feed-forward
 networks and need particular manual derivations and implementations when extende
d to other architectures. In this paper, we develop an automatic framework to en
able perturbation analysis on any neural network structures, by generalizing exi
sting LiRPA algorithms such as CROWN to operate on general computational graphs.
 The flexibility, differentiability and ease of use of our framework allow us to
 obtain state-of-the-art results on LiRPA based certified defense on fairly comp
licated networks like DenseNet, ResNeXt and Transformer that are not supported b
y prior works. Our framework also enables loss fusion, a technique that signific
antly reduces the computational complexity of LiRPA for certified defense. For t
he first time, we demonstrate LiRPA based certified defense on Tiny ImageNet and
 Downscaled ImageNet where previous approaches cannot scale to due to the relati
vely large number of classes. Our work also yields an open-source library for th
e community to apply LiRPA to areas beyond certified defense without much LiRPA
expertise, e.g., we create a neural network with a provably flat optimization la
ndscape by applying LiRPA to network parameters. Our open source library is avai
lable at https://github.com/KaidiXu/auto_LiRPA.
**********

Adaptation Properties Allow Identification of Optimized Neural Codes
Luke Rast, Jan Drugowitsch
The adaptation of neural codes to the statistics of their environment is well ca
ptured by efficient coding approaches. Here we solve an inverse problem: charact
erizing the objective and constraint functions that efficient codes appear to be
 optimal for, on the basis of how they adapt to different stimulus distributions
. We formulate a general efficient coding problem, with flexible objective and c
onstraint functions and minimal parametric assumptions. Solving special cases of
 this model, we provide solutions to broad classes of Fisher information-based e
fficient coding problems, generalizing a wide range of previous results. We show
 that different objective function types impose qualitatively different adaptati
on behaviors, while constraints enforce characteristic deviations from classic e
fficient coding signatures. Despite interaction between these effects, clear sig
natures emerge for both unconstrained optimization problems and information-maxi
mizing objective functions. Asking for a fixed-point of the neural code adaptati
on, we find an objective-independent characterization of constraints on the neur
al code. We use this result to propose an experimental paradigm that can charact
erize both the objective and constraint functions that an observed code appears
to be optimized for.
**********

Global Convergence and Variance Reduction for a Class of Nonconvex-Nonconcave Mi
nimax Problems
Junchi Yang, Negar Kiyavash, Niao He
Nonconvex minimax problems appear frequently in emerging machine learning applic
ations, such as generative adversarial networks and adversarial learning. Simple
 algorithms such as the gradient descent ascent (GDA) are the common practice fo
r solving these nonconvex games and receive lots of empirical success. Yet, it i
s known that these vanilla GDA algorithms with constant stepsize can potentially

diverge even in the convex setting. In this work, we show that for a subclass o
f nonconvex-nonconcave objectives satisfying a so-called two-sided Polyak-{\L}oj
asiewicz inequality, the alternating gradient descent ascent (AGDA) algorithm co
nverges globally at a linear rate and the stochastic AGDA achieves a sublinear r
ate. We further develop a variance reduced algorithm that attains a provably fas
ter rate than AGDA when the problem has the finite-sum structure.
**********

Model-Based Multi-Agent RL in Zero-Sum Markov Games with Near-Optimal Sample Com
plexity
Kaiqing Zhang, Sham Kakade, Tamer Basar, Lin 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.
**********

Conservative Q-Learning for Offline Reinforcement Learning
Aviral Kumar, Aurick Zhou, George Tucker, Sergey Levine
Effectively leveraging large, previously collected datasets in reinforcement lea
rn- ing (RL) is a key challenge for large-scale real-world applications. Offline
 RL algorithms promise to learn effective policies from previously-collected, st
atic datasets without further interaction. However, in practice, offline RL pres
ents a major challenge, and standard off-policy RL methods can fail due to overe
stimation of values induced by the distributional shift between the dataset and
the learned policy, especially when training on complex and multi-modal data dis
tributions. In this paper, we propose conservative Q-learning (CQL), which aims
to address these limitations by learning a conservative Q-function such that the
 expected value of a policy under this Q-function lower-bounds its true value. W
e theoretically show that CQL produces a lower bound on the value of the current
 policy and that it can be incorporated into a policy learning procedure with th
eoretical improvement guarantees. In practice, CQL augments the standard Bellman
 error objective with a simple Q-value regularizer which is straightforward to i
mplement on top of existing deep Q-learning and actor-critic implementations. On
 both discrete and continuous control domains, we show that CQL substantially ou
tperforms existing offline RL methods, often learning policies that attain 2-5 t
imes higher final return, especially when learning from complex and multi-modal
data distributions.
**********

Online Influence Maximization under Linear Threshold Model
Shuai Li, Fang Kong, Kejie Tang, Qizhi Li, Wei 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.
**********

Ensembling geophysical models with Bayesian Neural Networks
Ushnish Sengupta, Matt Amos, Scott Hosking, Carl Edward Rasmussen, Matthew Junip
er, Paul Young
Ensembles of geophysical models improve projection accuracy and express uncertai
nties. We develop a novel data-driven ensembling strategy for combining geophysi
cal models using Bayesian Neural Networks, which infers spatiotemporally varying
 model weights and bias while accounting for heteroscedastic uncertainties in th
e observations. This produces more accurate and uncertainty-aware projections wi
thout sacrificing interpretability. Applied to the prediction of total column oz
one from an ensemble of 15 chemistry-climate models, we find that the Bayesian n
eural network ensemble (BayNNE) outperforms existing ensembling methods, achievi
ng a 49.4% reduction in RMSE for temporal extrapolation, and a 67.4% reduction i
n RMSE for polar data voids, compared to a weighted mean. Uncertainty is also we
ll-characterized, with 90.6% of the data points in our extrapolation validation
dataset lying within 2 standard deviations and 98.5% within 3 standard deviation
s.

**********
Delving into the Cyclic Mechanism in Semi-supervised Video Object Segmentation
Yuxi Li, Ning Xu, Jinlong Peng, John See, Weiyao Lin
In this paper, we take attempt to incorporate the cyclic mechanism with the visi on task of semi-supervised video object segmentation. By resorting to the accura te reference mask of the first frame, we try to mitigate the error propagation p roblem in most of current video object segmentation pipelines. Firstly, we propo se a cyclic scheme for offline training of segmentation networks. Then, we exten d the offline pipeline to an online method by introducing a simple gradient corr ection module while keeping high efficiency as other offline methods. Finally we develop cycle effective receptive field (cycle-ERF) from gradient correction to provide a new perspective for analyzing object-specific regions of interests. W e conduct comprehensive experiments on benchmarks of DAVIS17 and Youtube-VOS, de monstrating that our introduced cyclic mechanism is helpful to boost the segment ation quality.
**********
Asymmetric Shapley values: incorporating causal knowledge into model-agnostic ex plainability
Christopher Frye, Colin Rowat, Ilya Feige
Explaining AI systems is fundamental both to the development of high performing models and to the trust placed in them by their users. The Shapley framework for explainability has strength in its general applicability combined with its prec ise, rigorous foundation: it provides a common, model-agnostic language for AI e xplainability and uniquely satisfies a set of intuitive mathematical axioms. How ever, Shapley values are too restrictive in one significant regard: they ignore all causal structure in the data. We introduce a less restrictive framework, Asy mmetric Shapley values (ASVs), which are rigorously founded on a set of axioms, applicable to any AI system, and can flexibly incorporate any causal structure k nown to be respected by the data. We demonstrate that ASVs can (i) improve model explanations by incorporating causal information, (ii) provide an unambiguous t est for unfair discrimination in model predictions, (iii) enable sequentially in cremental explanations in time-series models, and (iv) support feature-selection studies without the need for model retraining.
**********
Understanding Deep Architecture with Reasoning Layer
Xinshi Chen, Yufei Zhang, Christoph Reisinger, Le Song
Recently, there is a surge of interest in combining deep learning models with re asoning in order to handle more sophisticated learning tasks. In many cases, a r easoning task can be solved by an iterative algorithm. This algorithm is often u nrolled, truncated, and used as a specialized layer in the deep architecture, wh ich can be trained end-to-end with other neural components. Although such hybrid deep architectures have led to many empirical successes, theoretical understand ings of such architectures, especially the interplay between algorithm layers an d other neural layers, remains largely unexplored. In this paper, we take an ini tial step toward an understanding of such hybrid deep architectures by showing t hat properties of the algorithm layers, such as convergence, stability and sensi tivity, are intimately related to the approximation and generalization abilities of the end-to-end model. Furthermore, our analysis matches nicely with experime ntal observations under various conditions, suggesting that our theory can provi de useful guidelines for designing deep architectures with reasoning layers.
**********
Planning in Markov Decision Processes with Gap-Dependent Sample Complexity
Anders Jonsson, Emilie Kaufmann, Pierre Menard, Omar Darwiche Domingues, Edouard Leurent, Michal Valko
We propose MDP-GapE, a new trajectory-based Monte-Carlo Tree Search algorithm fo r planning in a Markov Decision Process in which transitions have a finite suppo rt. We prove an upper bound on the number of sampled trajectories needed for MDP -GapE to identify a near-optimal action with high probability. This problem-depe ndent result is expressed in terms of the sub-optimality gaps of the state-actio n pairs that are visited during exploration. Our experiments reveal that MDP-Gap

E is also effective in practice, in contrast with other algorithms with sample c
omplexity guarantees in the fixed-confidence setting, that are mostly theoretica
l.
**********

Provably Good Batch Off-Policy Reinforcement Learning Without Great Exploration
Yao Liu, Adith Swaminathan, Alekh Agarwal, Emma Brunskill
Batch reinforcement learning (RL) is important to apply RL algorithms to many hi
gh stakes tasks. Doing batch RL in a way that yields a reliable new policy in la
rge domains is challenging: a new decision policy may visit states and actions o
utside the support of the batch data, and function approximation and optimizatio
n with limited samples can further increase the potential of learning policies w
ith overly optimistic estimates of their future performance. Some recent approac
hes to address these concerns have shown promise, but can still be overly optimi
stic in their expected outcomes. Theoretical work that provides strong guarantee
s on the performance of the output policy relies on a strong concentrability ass
umption, which makes it unsuitable for cases where the ratio between state-actio
n distributions of behavior policy and some candidate policies is large. This is
 because, in the traditional analysis, the error bound scales up with this ratio
.
We show that using \emph{pessimistic value estimates} in the low-data regions in
 Bellman optimality and evaluation back-up can yield more adaptive and stronger
guarantees when the concentrability assumption does not hold. In certain setting
s, they can find the approximately best policy within the state-action space exp
lored by the batch data, without requiring a priori assumptions of concentrabili
ty. We highlight the necessity of our pessimistic update and the limitations of
previous algorithms and analyses by illustrative MDP examples and demonstrate an
 empirical comparison of our algorithm and other state-of-the-art batch RL basel
ines in standard benchmarks.
**********

Detection as Regression: Certified Object Detection with Median Smoothing
Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson,
Tom Goldstein
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Contextual Reserve Price Optimization in Auctions via Mixed Integer Programming
Joey Huchette, Haihao Lu, Hossein Esfandiari, Vahab Mirrokni
We study the problem of learning a linear model to set the reserve price in an a
uction, given contextual information, in order to maximize expected revenue from
 the seller side. First, we show that it is not possible to solve this problem i
n polynomial time unless the Exponential Time Hypothesis fails. Second, we prese
nt a strong mixed-integer programming (MIP) formulation for this problem, which
is capable of exactly modeling the nonconvex and discontinuous expected reward f
unction. Moreover, we show that this MIP formulation is ideal (i.e. the stronges
t possible formulation) for the revenue function of a single impression. Since i
t can be computationally expensive to exactly solve the MIP formulation in pract
ice, we also study the performance of its linear programming (LP) relaxation. Th
ough it may work well in practice, we show that, unfortunately, in the worst cas
e the optimal objective of the LP relaxation can be O(number of samples) times l
arger than the optimal objective of the true problem. Finally, we present comput
ational results, showcasing that the MIP formulation, along with its LP relaxati
on, are able to achieve superior in- and out-of-sample performance, as compared
to state-of-the-art algorithms on both real and synthetic datasets. More broadly
, we believe this work offers an indication of the strength of optimization meth
odologies like MIP to exactly model intrinsic discontinuities in machine learnin
g problems.
**********

ExpandNets: Linear Over-parameterization to Train Compact Convolutional Networks

Shuxuan Guo, Jose M. Alvarez, Mathieu Salzmann
We introduce an approach to training a given compact network. To this end, we le verage over-parameterization, which typically improves both neural network optim ization and generalization. Specifically, we propose to expand each linear layer of the compact network into multiple consecutive linear layers, without adding any nonlinearity. As such, the resulting expanded network, or ExpandNet, can be contracted back to the compact one algebraically at inference. In particular, we introduce two convolutional expansion strategies and demonstrate their benefits on several tasks, including image classification, object detection, and semanti c segmentation. As evidenced by our experiments, our approach outperforms both t raining the compact network from scratch and performing knowledge distillation f rom a teacher. Furthermore, our linear over-parameterization empirically reduces gradient confusion during training and improves the network generalization.
**********

FleXOR: Trainable Fractional Quantization
Dongsoo Lee, Se Jung Kwon, Byeongwook Kim, Yongkweon Jeon, Baeseong Park, Jeongi n Yun
Requests for name changes in the electronic proceedings will be accepted with 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 Implications of Local Correlation on Learning Some Deep Functions
Eran Malach, Shai Shalev-Shwartz
It is known that learning deep neural-networks is computationally hard in the wo rst-case. In fact, the proofs of such hardness results show that even weakly lea rning deep networks is hard. In other words, no efficient algorithm can find a p redictor that is slightly better than a random guess. However, we observe that o n natural distributions of images, small patches of the input image are corre- l ated to the target label, which implies that on such natural data, efficient wea k learning is trivial. While in the distribution-free setting, the celebrated bo osting results show that weak learning implies strong learning, in the distribut ion-specific setting this is not necessarily the case. We introduce a property o f distributions, denoted “local correlation”, which requires that small patches of the input image and of intermediate layers of the target function are correla ted to the target label. We empirically demonstrate that this property holds for the CIFAR and ImageNet data sets. The main technical results of the paper is pr oving that, for some classes of deep functions, weak learning implies efficient strong learning under the “local correlation” assumption.
**********

Learning to search efficiently for causally near-optimal treatments
Samuel Håkansson, Viktor Lindblom, Omer Gottesman, Fredrik D. Johansson
Finding an effective medical treatment often requires a search by trial and erro r. Making this search more efficient by minimizing the number of unnecessary tri als could lower both costs and patient suffering. We formalize this problem as l earning a policy for finding a near-optimal treatment in a minimum number of tri als using a causal inference framework. We give a model-based dynamic programmin g algorithm which learns from observational data while being robust to unmeasure d confounding. To reduce time complexity, we suggest a greedy algorithm which bo unds the near-optimality constraint. The methods are evaluated on synthetic and real-world healthcare data and compared to model-free reinforcement learning. We find that our methods compare favorably to the model-free baseline while offeri ng a more transparent trade-off between search time and treatment efficacy.
**********

A Game Theoretic Analysis of Additive Adversarial Attacks and Defenses
Ambar Pal, Rene Vidal
Research in adversarial learning follows a cat and mouse game between attackers and defenders where attacks are proposed, they are mitigated by new defenses, an d subsequently new attacks are proposed that break earlier defenses, and so on. However, it has remained unclear as to whether there are conditions under which

no better attacks or defenses can be proposed. In this paper, we propose a game-theoretic framework for studying attacks and defenses which exist in equilibrium. Under a locally linear decision boundary model for the underlying binary classifier, we prove that the Fast Gradient Method attack and a Randomized Smoothing defense form a Nash Equilibrium. We then show how this equilibrium defense can be approximated given finitely many samples from a data-generating distribution, and derive a generalization bound for the performance of our approximation.
**********
Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts
Bertrand Charpentier, Daniel Zügner, Stephan Günnemann
Accurate estimation of aleatoric and epistemic uncertainty is crucial to build safe and reliable systems. Traditional approaches, such as dropout and ensemble methods, estimate uncertainty by sampling probability predictions from different submodels, which leads to slow uncertainty estimation at inference time. Recent works address this drawback by directly predicting parameters of prior distributions over the probability predictions with a neural network. While this approach has demonstrated accurate uncertainty estimation, it requires defining arbitrary target parameters for in-distribution data and makes the unrealistic assumption that out-of-distribution (OOD) data is known at training time.
**********
Recurrent Quantum Neural Networks
Johannes Bausch
Recurrent neural networks are the foundation of many sequence-to-sequence models in machine learning, such as machine translation and speech synthesis. With applied quantum computing in its infancy, there already exist quantum machine learning models such as variational quantum eigensolvers which have been used e.g. in the context of energy minimization tasks. Yet, to date, no viable recurrent quantum network has been proposed.
**********
No-Regret Learning and Mixed Nash Equilibria: They Do Not Mix
Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, Georgios Piliouras
Understanding the behavior of no-regret dynamics in general N-player games is a fundamental question in online learning and game theory. A folk result in the field states that, in finite games, the empirical frequency of play under no-regret
learning converges to the game’s set of coarse correlated equilibria. By contrast,
our understanding of how the day-to-day behavior of the dynamics correlates to the game’s Nash equilibria is much more limited, and only partial results are known
for certain classes of games (such as zero-sum or congestion games). In this paper, we study the dynamics of follow the regularized leader (FTRL), arguably the most well-studied class of no-regret dynamics, and we establish a sweeping negative result showing that the notion of mixed Nash equilibrium is antithetical to no-regret learning. Specifically, we show that any Nash equilibrium which is not strict (in that every player has a unique best response) cannot be stable and attracting under the dynamics of FTRL. This result has significant implications for predicting the outcome of a learning process as it shows unequivocally that only strict (and hence, pure) Nash equilibria can emerge as stable limit points thereof.
**********
A Unifying View of Optimism in Episodic Reinforcement Learning
Gergely Neu, Ciara Pike-Burke
The principle of ``optimism in the face of uncertainty'' underpins many theoretically successful reinforcement learning algorithms. In this paper we provide a general framework for designing, analyzing and implementing such algorithms in the episodic reinforcement learning problem. This framework is built upon Lagrangian duality, and demonstrates that every model-optimistic algorithm that construc

ts an optimistic MDP has an equivalent representation as a value-optimistic dyna
mic programming algorithm. Typically, it was thought that these two classes of a
lgorithms were distinct, with model-optimistic algorithms benefiting from a clea
ner probabilistic analysis while value-optimistic algorithms are easier to imple
ment and thus more practical. With the framework developed in this paper, we sho
w that it is possible to get the best of both worlds by providing a class of alg
orithms which have a computationally efficient dynamic-programming implementatio
n and also a simple probabilistic analysis. Besides being able to capture many e
xisting algorithms in the tabular setting, our framework can also address large-
scale problems under realizable function approximation, where it enables a simpl
e model-based analysis of some recently proposed methods.
**********

Continuous Submodular Maximization: Beyond DR-Submodularity
Moran Feldman, 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.
**********

An Asymptotically Optimal Primal-Dual Incremental Algorithm for Contextual Linea
r Bandits
Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, Alessandro Lazaric
In the contextual linear bandit setting, algorithms built on the optimism princi
ple fail to exploit the structure of the problem and have been shown to be asymp
totically suboptimal. In this paper, we follow recent approaches of deriving asy
mptotically optimal algorithms from problem-dependent regret lower bounds and we
 introduce a novel algorithm improving over the state-of-the-art along multiple
dimensions. We build on a reformulation of the lower bound, where context distri
bution and exploration policy are decoupled, and we obtain an algorithm robust t
o unbalanced context distributions. Then, using an incremental primal-dual appro
ach to solve the Lagrangian relaxation of the lower bound, we obtain a scalable
and computationally efficient algorithm. Finally, we remove forced exploration a
nd build on confidence intervals of the optimization problem to encourage a mini
mum level of exploration that is better adapted to the problem structure. We dem
onstrate the asymptotic optimality of our algorithm, while providing both proble
m-dependent and worst-case finite-time regret guarantees. Our bounds scale with
the logarithm of the number of arms, thus avoiding the linear dependence common
in all related prior works. Notably, we establish minimax optimality for any lea
rning horizon in the special case of non-contextual linear bandits. Finally, we
verify that our algorithm obtains better empirical performance than state-of-the
-art baselines.
**********

Assessing SATNet's Ability to Solve the Symbol Grounding Problem
Oscar Chang, Lampros Flokas, Hod Lipson, Michael Spranger
SATNet is an award-winning MAXSAT solver that can be used to infer logical rules
 and integrated as a differentiable layer in a deep neural network. It had been
shown to solve Sudoku puzzles visually from examples of puzzle digit images, and
 was heralded as an impressive achievement towards the longstanding AI goal of c
ombining pattern recognition with logical reasoning. In this paper, we clarify S
ATNet's capabilities by showing that in the absence of intermediate labels that
identify individual Sudoku digit images with their logical representations, SATN
et completely fails at visual Sudoku (0% test accuracy). More generally, the fai
lure can be pinpointed to its inability to learn to assign symbols to perceptual
 phenomena, also known as the symbol grounding problem, which has long been thou
ght to be a prerequisite for intelligent agents to perform real-world logical re
asoning. We propose an MNIST based test as an easy instance of the symbol ground
ing problem that can serve as a sanity check for differentiable symbolic solvers
 in general. Naive applications of SATNet on this test lead to performance worse
 than that of models without logical reasoning capabilities. We report on the ca
uses of SATNet’s failure and how to prevent them.

**********
A Bayesian Nonparametrics View into Deep Representations
Micha■ Jamro■, Marcin Kurdziel, Mateusz Opala
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Similarity between the Laplace and Neural Tangent Kernels
Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, Basri Ronen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 causal view of compositional zero-shot recognition
Yuval Atzmon, Felix Kreuk, Uri Shalit, Gal Chechik
People easily recognize new visual categories that are new combinations of known components. This compositional generalization capacity is critical for learning in real-world domains like vision and language because the long tail of new combinations dominates the distribution. Unfortunately, learning systems struggle with compositional generalization because they often build on features that are correlated with class labels even if they are not "essential" for the class. This leads to consistent misclassification of samples from a new distribution, like new combinations of known components.
**********
HiPPO: Recurrent Memory with Optimal Polynomial Projections
Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, Christopher Ré
A central problem in learning from sequential data is representing cumulative history in an incremental fashion as more data is processed. We introduce a general framework (HiPPO) for the online compression of continuous signals and discrete time series by projection onto polynomial bases. Given a measure that specifies the importance of each time step in the past, HiPPO produces an optimal solution to a natural online function approximation problem. As special cases, our framework yields a short derivation of the recent Legendre Memory Unit (LMU) from first principles, and generalizes the ubiquitous gating mechanism of recurrent neural networks such as GRUs. This formal framework yields a new memory update mechanism (HiPPO-LegS) that scales through time to remember all history, avoiding priors on the timescale. HiPPO-LegS enjoys the theoretical benefits of timescale robustness, fast updates, and bounded gradients. By incorporating the memory dynamics into recurrent neural networks, HiPPO RNNs can empirically capture complex temporal dependencies. On the benchmark permuted MNIST dataset, HiPPO-LegS sets a new state-of-the-art accuracy of 98.3%. Finally, on a novel trajectory classification task testing robustness to out-of-distribution timescales and missing data, HiPPO-LegS outperforms RNN and neural ODE baselines by 25-40% accuracy.
**********
Auto Learning Attention
Benteng Ma, Jing Zhang, Yong Xia, Dacheng Tao
Attention modules have been demonstrated effective in strengthening the representation ability of a neural network via reweighting spatial or channel features or stacking both operations sequentially. However, designing the structures of different attention operations requires a bulk of computation and extensive expertise. In this paper, we devise an Auto Learning Attention (AutoLA) method, which is the first attempt on automatic attention design. Specifically, we define a novel attention module named high order group attention (HOGA) as a directed acyclic graph (DAG) where each group represents a node, and each edge represents an operation of heterogeneous attentions. A typical HOGA architecture can be searched automatically via the differential AutoLA method within 1 GPU day using the ResNet-20 backbone on CIFAR10. Further, the searched attention module can generalize to various backbones as a plug-and-play component and outperforms popular man

ually designed channel and spatial attentions for many vision tasks, including i
mage classification on CIFAR100 and ImageNet, object detection and human keypoin
t detection on COCO dataset. The code will be released.
**********

CASTLE: Regularization via Auxiliary Causal Graph Discovery
Trent Kyono, Yao Zhang, Mihaela van der Schaar
Regularization improves generalization of supervised models to out-of-sample dat
a. Prior works have shown that prediction in the causal direction (effect from c
ause) results in lower testing error than the anti-causal direction. However, ex
isting regularization methods are agnostic of causality. We introduce Causal Str
ucture Learning (CASTLE) regularization and propose to regularize a neural netwo
rk by jointly learning the causal relationships between variables. CASTLE learns
 the causal directed acyclical graph (DAG) as an adjacency matrix embedded in th
e neural network's input layers, thereby facilitating the discovery of optimal p
redictors. Furthermore, CASTLE efficiently reconstructs only the features in the
 causal DAG that have a causal neighbor, whereas reconstruction-based regularize
rs suboptimally reconstruct all input features. We provide a theoretical general
ization bound for our approach and conduct experiments on a plethora of syntheti
c and real publicly available datasets demonstrating that CASTLE consistently le
ads to better out-of-sample predictions as compared to other popular benchmark r
egularizers.
**********

Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Cau
sal Effect
Kaihua Tang, Jianqiang Huang, Hanwang Zhang
As the class size grows, maintaining a balanced dataset across many classes is c
hallenging because the data are long-tailed in nature; it is even impossible whe
n the sample-of-interest co-exists with each other in one collectable unit, e.g.
, multiple visual instances in one image. Therefore, long-tailed classification
is the key to deep learning at scale. However, existing methods are mainly based
 on re-weighting/re-sampling heuristics that lack a fundamental theory. In this
paper, we establish a causal inference framework, which not only unravels the wh
ys of previous methods, but also derives a new principled solution. Specifically
, our theory shows that the SGD momentum is essentially a confounder in long-tai
led classification. On one hand, it has a harmful causal effect that misleads th
e tail prediction biased towards the head. On the other hand, its induced mediat
ion also benefits the representation learning and head prediction. Our framework
 elegantly disentangles the paradoxical effects of the momentum, by pursuing the
 direct causal effect caused by an input sample. In particular, we use causal in
tervention in training, and counterfactual reasoning in inference, to remove the
 bad'' while keep thegood''. We achieve new state-of-the-arts on three long-tail
ed visual recognition benchmarks: Long-tailed CIFAR-10/-100, ImageNet-LT for ima
ge classification and LVIS for instance segmentation.
**********

Explainable Voting
Dominik Peters, Ariel D. Procaccia, Alexandros Psomas, Zixin 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Deep Archimedean Copulas
Chun Kai Ling, Fei Fang, J. Zico Kolter
A central problem in machine learning and statistics is to model joint densities
 of random variables from data. Copulas are joint cumulative distribution functi
ons with uniform marginal distributions and are used to capture interdependencie
s in isolation from marginals. Copulas are widely used within statistics, but ha
ve not gained traction in the context of modern deep learning. In this paper, we
 introduce ACNet, a novel differentiable neural network architecture that enforc
es structural properties and enables one to learn an important class of copulas-

-Archimedean Copulas. Unlike Generative Adversarial Networks, Variational Autoen
coders, or Normalizing Flow methods, which learn either densities or the generat
ive process directly, ACNet learns a generator of the copula, which implicitly d
efines the cumulative distribution function of a joint distribution. We give a p
robabilistic interpretation of the network parameters of ACNet and use this to d
erive a simple but efficient sampling algorithm for the learned copula. Our expe
riments show that ACNet is able to both approximate common Archimedean Copulas a
nd generate new copulas which may provide better fits to data.
**********

Re-Examining Linear Embeddings for High-Dimensional Bayesian Optimization
Ben Letham, Roberto Calandra, Akshara Rai, Eytan Bakshy
Bayesian optimization (BO) is a popular approach to optimize expensive-to-evalua
te black-box functions. A significant challenge in BO is to scale to high-dimens
ional parameter spaces while retaining sample efficiency. A solution considered
in existing literature is to embed the high-dimensional space in a lower-dimensi
onal manifold, often via a random linear embedding. In this paper, we identify s
everal crucial issues and misconceptions about the use of linear embeddings for
BO. We study the properties of linear embeddings from the literature and show th
at some of the design choices in current approaches adversely impact their perfo
rmance. We show empirically that properly addressing these issues significantly
improves the efficacy of linear embeddings for BO on a range of problems, includ
ing learning a gait policy for robot locomotion.
**********

UnModNet: Learning to Unwrap a Modulo Image for High Dynamic Range Imaging
Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, Boxin Shi
A conventional camera often suffers from over- or under-exposure when recording
a real-world scene with a very high dynamic range (HDR). In contrast, a modulo c
amera with a Markov random field (MRF) based unwrapping algorithm can theoretica
lly accomplish unbounded dynamic range but shows degenerate performances when th
ere are modulus-intensity ambiguity, strong local contrast, and color misalignme
nt. In this paper, we reformulate the modulo image unwrapping problem into a ser
ies of binary labeling problems and propose a modulo edge-aware model, named as
UnModNet, to iteratively estimate the binary rollover masks of the modulo image
for unwrapping. Experimental results show that our approach can generate 12-bit
HDR images from 8-bit modulo images reliably, and runs much faster than the prev
ious MRF-based algorithm thanks to the GPU acceleration.
**********

Thunder: a Fast Coordinate Selection Solver for  Sparse Learning
Shaogang Ren, Weijie Zhao, Ping Li
L1 regularization has been broadly employed to pursue model sparsity. Despite th
e non-smoothness, people have developed efficient algorithms by leveraging the s
parsity and convexity of the problems. In this paper, we propose a novel active
incremental approach to further improve the efficiency of the solvers. We show t
hat our method performs well even when the existing methods fail due to the low
sparseness or high solution accuracy request. Theoretical analysis and experimen
tal results on synthetic and real-world data sets validate the advantages of the
 method.
**********

Neural Networks Fail to Learn Periodic Functions and How to Fix It
Liu Ziyin, Tilman Hartwig, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Distribution Matching for Crowd Counting
Boyu Wang, Huidong Liu, Dimitris Samaras, Minh Hoai Nguyen
In crowd counting, each training image contains multiple people, where each pers
on is annotated by a dot. Existing crowd counting methods need to use a Gaussian
 to smooth each annotated dot or to estimate the likelihood of every pixel given

the annotated point. In this paper, we show that imposing Gaussians to annotations hurts generalization performance. Instead, we propose to use Distribution Matching for crowd COUNTing (DM-Count). In DM-Count, we use Optimal Transport (OT) to measure the similarity between the normalized predicted density map and the normalized ground truth density map. To stabilize OT computation, we include a Total Variation loss in our model. We show that the generalization error bound of DM-Count is tighter than that of the Gaussian smoothed methods. In terms of Mean Absolute Error, DM-Count outperforms the previous state-of-the-art methods by a large margin on two large-scale counting datasets, UCF-QNRF and NWPU, and achieves the state-of-the-art results on the ShanghaiTech and UCF-CC50 datasets. DM-Count reduced the error of the state-of-the-art published result by approximately 16%. Code is available at https://github.com/cvlab-stonybrook/DM-Count.

**********

Correspondence learning via linearly-invariant embedding

Riccardo Marin, Marie-Julie Rakotosaona, Simone Melzi, Maks Ovsjanikov

In this paper, we propose a fully differentiable pipeline for estimating accurate dense correspondences between 3D point clouds. The proposed pipeline is an extension and a generalization of the functional maps framework. However, instead of using the Laplace-Beltrami eigenfunctions as done in virtually all previous works in this domain, we demonstrate that learning the basis from data can both improve robustness and lead to better accuracy in challenging settings. We interpret the basis as a learned embedding into a higher dimensional space. Following the functional map paradigm the optimal transformation in this embedding space must be linear and we propose a separate architecture aimed at estimating the transformation by learning optimal descriptor functions. This leads to the first end-to-end trainable functional map-based correspondence approach in which both the basis and the descriptors are learned from data. Interestingly, we also observe that learning a canonical embedding leads to worse results, suggesting that leaving an extra linear degree of freedom to the embedding network gives it more robustness, thereby also shedding light onto the success of previous methods. Finally, we demonstrate that our approach achieves state-of-the-art results in challenging non-rigid 3D point cloud correspondence applications.

**********

Learning to Dispatch for Job Shop Scheduling via Deep Reinforcement Learning

Cong Zhang, Wen Song, Zhiguang Cao, Jie Zhang, Puay Siew Tan, Xu Chi

Priority dispatching rule (PDR) is widely used for solving real-world Job-shop scheduling problem (JSSP). However, the design of effective PDRs is a tedious task, requiring a myriad of specialized knowledge and often delivering limited performance. In this paper, we propose to automatically learn PDRs via an end-to-end deep reinforcement learning agent. We exploit the disjunctive graph representation of JSSP, and propose a Graph Neural Network based scheme to embed the states encountered during solving. The resulting policy network is size-agnostic, effectively enabling generalization on large-scale instances. Experiments show that the agent can learn high-quality PDRs from scratch with elementary raw features, and demonstrates strong performance against the best existing PDRs. The learned policies also perform well on much larger instances that are unseen in training.

**********

On Adaptive Attacks to Adversarial Example Defenses

Florian Tramer, Nicholas Carlini, Wieland Brendel, Aleksander Madry

Adaptive attacks have (rightfully) become the de facto standard for evaluating defenses to adversarial examples. We find, however, that typical adaptive evaluations are incomplete.

We demonstrate that 13 defenses recently published at ICLR, ICML and NeurIPS---and which illustrate a diverse set of defense strategies---can be circumvented despite attempting to perform evaluations using adaptive attacks.

**********

Sinkhorn Natural Gradient for Generative Models

Zebang Shen, Zhenfu Wang, Alejandro Ribeiro, Hamed Hassani

Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Online Sinkhorn: Optimal Transport distances from sample streams
Arthur Mensch, Gabriel Peyré
Optimal Transport (OT) distances are now routinely used as loss functions in ML tasks. Yet, computing OT distances between arbitrary (i.e. not necessarily discr ete) probability distributions remains an open problem. This paper introduces a new online estimator of entropy-regularized OT distances between two such arbitr ary distributions. It uses streams of samples from both distributions to iterati vely enrich a non-parametric representation of the transportation plan. Compared to the classic Sinkhorn algorithm, our method leverages new samples at each ite ration, which enables a consistent estimation of the true regularized OT distanc e. We provide a theoretical analysis of the convergence of the online Sinkhorn a lgorithm, showing a nearly-1/n asymptotic sample complexity for the iterate sequ ence. We validate our method on synthetic 1-d to 10-d data and on real 3-d shape data.
**********

Ultrahyperbolic Representation Learning
Marc Law, Jos Stam
In machine learning, data is usually represented in a (flat) Euclidean space whe re distances between points are along straight lines. Researchers have recently considered more exotic (non-Euclidean) Riemannian manifolds such as hyperbolic s pace which is well suited for tree-like data. In this paper, we propose a repres entation living on a pseudo-Riemannian manifold of constant nonzero curvature. I t is a generalization of hyperbolic and spherical geometries where the non-degen erate metric tensor need not be positive definite. We provide the necessary lear ning tools in this geometry and extend gradient method optimization techniques. More specifically, we provide closed-form expressions for distances via geodesic s and define a descent direction to minimize some objective function. Our novel framework is applied to graph representations.
**********

Locally-Adaptive Nonparametric Online Learning
Ilja Kuzborskij, Nicolò Cesa-Bianchi
One of the main strengths of online algorithms is their ability to adapt to arbi trary data sequences. This is especially important in nonparametric settings, wh ere performance is measured against rich classes of comparator functions that ar e able to fit complex environments. Although such hard comparators and complex e nvironments may exhibit local regularities, efficient algorithms, which can prov ably take advantage of these local patterns, are hardly known. We fill this gap by introducing efficient online algorithms (based on a single versatile master a lgorithm) each adapting to one of the following regularities: (i) local Lipschit zness of the competitor function, (ii) local metric dimension of the instance se quence, (iii) local performance of the predictor across different regions of the instance space. Extending previous approaches, we design algorithms that dynami cally grow hierarchical ε-nets on the instance space whose prunings correspond t o different “locality profiles” for the problem at hand. Using a technique based on tree experts, we simultaneously and efficiently compete against all such pru nings, and prove regret bounds each scaling with a quantity associated with a di fferent type of local regularity. When competing against “simple” locality profi les, our technique delivers regret bounds that are significantly better than tho se proven using the previous approach. On the other hand, the time dependence of our bounds is not worse than that obtained by ignoring any local regularities.
**********

Compositional Generalization via Neural-Symbolic Stack Machines
Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, Denny Zhou
Despite achieving tremendous success, existing deep learning models have exposed limitations in compositional generalization, the capability to learn compositio nal rules and apply them to unseen cases in a systematic manner. To tackle this

issue, we propose the Neural-Symbolic Stack Machine (NeSS). It contains a neural network to generate traces, which are then executed by a symbolic stack machine enhanced with sequence manipulation operations. NeSS combines the expressive power of neural sequence models with the recursion supported by the symbolic stack machine. Without training supervision on execution traces, NeSS achieves 100% generalization performance in four domains: the SCAN benchmark of language-driven navigation tasks, the task of few-shot learning of compositional instructions, the compositional machine translation benchmark, and context-free grammar parsing tasks.

**********

Graphon Neural Networks and the Transferability of Graph Neural Networks

Luana Ruiz, Luiz Chamon, Alejandro Ribeiro

Graph neural networks (GNNs) rely on graph convolutions to extract local features from network data. These graph convolutions combine information from adjacent nodes using coefficients that are shared across all nodes.

Since these coefficients are shared and do not depend on the graph, one can envision using the same coefficients to define a GNN on another graph. This motivates analyzing the transferability of GNNs across graphs.

In this paper we introduce graphon NNs as limit objects of GNNs and prove a bound on the difference between the output of a GNN and its limit graphon-NN. This bound vanishes with growing number of nodes if the graph convolutional filters are bandlimited in the graph spectral domain. This result establishes a tradeoff between discriminability and transferability of GNNs.

**********

Unreasonable Effectiveness of Greedy Algorithms in Multi-Armed Bandit with Many Arms

Mohsen Bayati, Nima Hamidi, Ramesh Johari, Khashayar Khosravi

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

**********

Gamma-Models: Generative Temporal Difference Learning for Infinite-Horizon Prediction

Michael Janner, Igor Mordatch, Sergey Levine

We introduce the gamma-model, a predictive model of environment dynamics with an infinite, probabilistic horizon. Replacing standard single-step models with gamma-models leads to generalizations of the procedures that form the foundation of model-based control, including the model rollout and model-based value estimation. The gamma-model, trained with a generative reinterpretation of temporal difference learning, is a natural continuous analogue of the successor representation and a hybrid between model-free and model-based mechanisms. Like a value function, it contains information about the long-term future; like a standard predictive model, it is independent of task reward. We instantiate the gamma-model as both a generative adversarial network and normalizing flow, discuss how its training reflects an inescapable tradeoff between training-time and testing-time compounding errors, and empirically investigate its utility for prediction and control.

**********

Deep Transformers with Latent Depth

Xian Li, Asa Cooper Stickland, Yuqing Tang, Xiang Kong

The Transformer model has achieved state-of-the-art performance in many sequence modeling tasks. However, how to leverage model capacity with large or variable depths is still an open challenge. We present a probabilistic framework to automatically learn which layer(s) to use by learning the posterior distributions of layer selection. As an extension of this framework, we propose a novel method to train one shared Transformer network for multilingual machine translation with different layer selection posteriors for each language pair. The proposed method alleviates the vanishing gradient issue and enables stable training of deep Transformers (e.g. 100 layers). We evaluate on WMT English-German machine translati

on and masked language modeling tasks, where our method outperforms existing app
roaches for training deeper Transformers. Experiments on multilingual machine tr
anslation demonstrate that this approach can effectively leverage increased mode
l capacity and bring universal improvement for both many-to-one and one-to-many
translation with diverse language pairs.
**********

Neural Mesh Flow: 3D Manifold Mesh Generation via Diffeomorphic  Flows
Kunal Gupta, Manmohan Chandraker
Meshes are important representations of physical 3D entities in the virtual worl
d. Applications like rendering, simulations and 3D printing require meshes to be
 manifold so that they can interact with the world like the real objects they re
present.  Prior methods generate meshes with great geometric accuracy but poor m
anifoldness. In this work, we propose NeuralMeshFlow (NMF) to generate two-manif
old meshes for genus-0 shapes. Specifically, NMF is a shape auto-encoder consist
ing of several Neural Ordinary Differential Equation (NODE)(1) blocks that learn
 accurate mesh geometry by progressively deforming a spherical mesh. Training NM
F is simpler compared to state-of-the-art methods since it does not require any
explicit mesh-based regularization. Our experiments demonstrate that NMF facilit
ates several applications such as single-view mesh reconstruction, global shape
parameterization, texture mapping, shape deformation and correspondence. Importa
ntly, we demonstrate that manifold meshes generated using NMF are better-suited
for physically-based rendering and simulation compared to prior works.
**********

Statistical control for spatio-temporal MEG/EEG source imaging with desparsified
 mutli-task Lasso
Jerome-Alexis Chevalier, Joseph Salmon, Alexandre Gramfort, Bertrand Thirion
Detecting where and when brain regions activate in a cognitive task or in a give
n clinical condition is the promise of non-invasive techniques like magnetoencep
halography (MEG) or electroencephalography (EEG). This problem, referred to as s
ource localization, or source imaging, poses however a high-dimensional statisti
cal inference challenge. While sparsity promoting regularizations have been prop
osed to address the regression problem, it remains unclear how to ensure statist
ical control of false detections in this setting. Moreover, MEG/EEG source imagi
ng requires to work with spatio-temporal data and autocorrelated noise. To deal
with this, we adapt the desparsified Lasso estimator ---an estimator tailored fo
r high dimensional linear model that asymptotically follows a Gaussian distribut
ion under sparsity and moderate feature correlation assumptions--- to temporal d
ata corrupted with autocorrelated noise. We call it the desparsified multi-task
Lasso (d-MTLasso). We combine d-MTLasso with spatially constrained clustering to
 reduce data dimension and with ensembling to mitigate the arbitrary choice of c
lustering; the resulting estimator is called ensemble of clustered desparsified
multi-task Lasso (ecd-MTLasso). With respect to the current procedures, the two
advantages of ecd-MTLasso are that i)it offers statistical guarantees and ii)it
allows to trade spatial specificity for sensitivity, leading to a powerful adapt
ive method. Extensive simulations on realistic head geometries, as well as empir
ical results on various MEG datasets, demonstrate the high recovery performance
of ecd-MTLasso and its primary practical benefit: offer a statistically principl
ed way to threshold MEG/EEG source maps.
**********

A Scalable MIP-based Method for Learning Optimal Multivariate Decision Trees
Haoran Zhu, Pavankumar Murali, Dzung Phan, Lam Nguyen, Jayant Kalagnanam
Several recent publications report advances in training optimal decision trees (
ODTs) using mixed-integer programs (MIPs), due to algorithmic advances in intege
r programming and a growing interest in addressing the inherent suboptimality of
 heuristic approaches such as CART. In this paper, we propose a novel MIP formul
ation, based on 1-norm support vector machine model, to train a binary oblique O
DT for classification problems. We further present techniques, such as cutting p
lanes, to tighten its linear relaxation, to improve run times to reach optimalit
y. Using 36 datasets from the University of California Irvine Machine Learning R
epository, we demonstrate that our training approach outperforms its counterpart

s from literature in terms of out-of-sample performance (around 10% improvement in mean out-of-sample testing accuracy). Towards our goal of developing a scalable framework to train multivariate ODT on large datasets, we propose a new linear programming based data selection method to choose a subset of the data, and use it to train a decision tree through our proposed MIP model. We conclude this paper with extensive numerical testing results, that showcase the generalization performance of our new MIP formulation, and the improvement in mean out-of-sample accuracy on large datasets.
**********
Efficient Exact Verification of Binarized Neural Networks
Kai Jia, Martin Rinard
Concerned with the reliability of neural networks, researchers have developed verification techniques to prove their robustness. Most verifiers work with real-valued networks. Unfortunately, the exact (complete and sound) verifiers face scalability challenges and provide no correctness guarantees due to floating point errors. We argue that Binarized Neural Networks (BNNs) provide comparable robustness and allow exact and significantly more efficient verification. We present a new system, EEV, for efficient and exact verification of BNNs. EEV consists of two parts: (i) a novel SAT solver that speeds up BNN verification by natively handling the reified cardinality constraints arising in BNN encodings; and (ii) strategies to train solver-friendly robust BNNs by inducing balanced layer-wise sparsity and low cardinality bounds, and adaptively cancelling the gradients. We demonstrate the effectiveness of EEV by presenting the first exact verification results for L-inf-bounded adversarial robustness of nontrivial convolutional BNNs on the MNIST and CIFAR10 datasets. Compared to exact verification of real-valued networks of the same architectures on the same tasks, EEV verifies BNNs hundreds to thousands of times faster, while delivering comparable verifiable accuracy in most cases.
**********
Ultra-Low Precision 4-bit Training of Deep Neural Networks
Xiao Sun, Naigang Wang, Chia-Yu Chen, Jiamin Ni, Ankur Agrawal, Xiaodong Cui, Swagath Venkataramani, Kaoutar El Maghraoui, Vijayalakshmi (Viji) Srinivasan, Kailash Gopalakrishnan
In this paper, we propose a number of novel techniques and numerical representation formats that enable, for the very first time, the precision of training systems to be aggressively scaled from 8-bits to 4-bits. To enable this advance, we explore a novel adaptive Gradient Scaling technique (Gradscale) that addresses the challenges of insufficient range and resolution in quantized gradients as well as explores the impact of quantization errors observed during model training. We theoretically analyze the role of bias in gradient quantization and propose solutions that mitigate the impact of this bias on model convergence. Finally, we examine our techniques on a spectrum of deep learning models in computer vision, speech, and NLP. In combination with previously proposed solutions for 4-bit quantization of weight and activation tensors, 4-bit training shows a non-significant loss in accuracy across application domains while enabling significant hardware acceleration (> 7X over state-of-the-art FP16 systems).
**********
Bridging the Gap between Sample-based and One-shot Neural Architecture Search with BONAS
Han Shi, Renjie Pi, Hang Xu, Zhenguo Li, James Kwok, Tong Zhang
Neural Architecture Search (NAS) has shown great potentials in finding better neural network designs. Sample-based NAS is the most reliable approach which aims at exploring the search space and evaluating the most promising architectures. However, it is computationally very costly.  As a remedy, the one-shot approach has emerged as a popular technique for accelerating NAS using weight-sharing. However, due to the weight-sharing of vastly different networks, the one-shot approach is less reliable than the sample-based approach. In this work, we propose BONAS (Bayesian Optimized Neural Architecture Search), a sample-based NAS framework which is accelerated using weight-sharing to evaluate multiple related architectures simultaneously. Specifically, we apply Graph Convolutional Network predic

tor as a surrogate model for Bayesian Optimization to select multiple related candidate models in each iteration. We then apply weight-sharing to train multiple candidate models simultaneously. This approach not only accelerates the traditional sample-based approach significantly, but also keeps its reliability. This is because weight-sharing among related architectures are more reliable than those in the one-shot approach. Extensive experiments are conducted to verify the effectiveness of our method over many competing algorithms.

**********

On Numerosity of Deep Neural Networks

Xi Zhang, Xiaolin Wu

Recently, a provocative claim was published that number sense spontaneously emerges in a deep neural network trained merely for visual object recognition. This has, if true, far reaching significance to the fields of machine learning and cognitive science alike. In this paper, we prove the above claim to be unfortunately incorrect. The statistical analysis to support the claim is flawed in that the sample set used to identify number-aware neurons is too small, compared to the huge number of neurons in the object recognition network. By this flawed analysis one could mistakenly identify number-sensing neurons in any randomly initialized deep neural networks that are not trained at all. With the above critique we ask the question what if a deep convolutional neural network is carefully trained for numerosity? Our findings are mixed. Even after being trained with number-depicting images, the deep learning approach still has difficulties to acquire the abstract concept of numbers, a cognitive task that preschoolers perform with ease. But on the other hand, we do find some encouraging evidences suggesting that deep neural networks are more robust to distribution shift for small numbers than for large numbers.

**********

Outlier Robust Mean Estimation with Subgaussian Rates via Stability

Ilias Diakonikolas, Daniel M. Kane, Ankit Pensia

We study the problem of outlier robust high-dimensional mean estimation under a bounded covariance assumption, and more broadly under bounded low-degree moment assumptions. We consider a standard stability condition from the recent robust statistics literature and prove that, except with exponentially small failure probability, there exists a large fraction of the inliers satisfying this condition. As a corollary, it follows that a number of recently developed algorithms for robust mean estimation, including iterative filtering and non-convex gradient descent, give optimal error estimators with (near-)subgaussian rates. Previous analyses of these algorithms gave significantly suboptimal rates. As a corollary of our approach, we obtain the first computationally efficient algorithm for outlier robust mean estimation with subgaussian rates under a bounded covariance assumption.

**********

Self-Supervised Relationship Probing

Jiuxiang Gu, Jason Kuen, Shafiq Joty, Jianfei Cai, Vlad Morariu, Handong Zhao, Tong Sun

Structured representations of images that model visual relationships are beneficial for many vision and vision-language applications. However, current human-annotated visual relationship datasets suffer from the long-tailed predicate distribution problem which limits the potential of visual relationship models. In this work, we introduce a self-supervised method that implicitly learns the visual relationships without relying on any ground-truth visual relationship annotations. Our method relies on 1) intra- and inter-modality encodings to respectively model relationships within each modality separately and jointly, and 2) relationship probing, which seeks to discover the graph structure within each modality. By leveraging masked language modeling, contrastive learning, and dependency tree distances for self-supervision, our method learns better object features as well as implicit visual relationships. We verify the effectiveness of our proposed method on various vision-language tasks that benefit from improved visual relationship understanding.

**********

Information Theoretic Counterfactual Learning from Missing-Not-At-Random Feedbac
k
Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan Kuruoglu, Yefeng Zheng
Counterfactual learning for dealing with missing-not-at-random data (MNAR) is an
 intriguing topic in the recommendation literature, since MNAR data are ubiquito
us in modern recommender systems. Instead, missing-at-random (MAR) data, namely
randomized controlled trials (RCTs), are usually required by most previous count
erfactual learning methods. However, the execution of RCTs is extraordinarily ex
pensive in practice. To circumvent the use of RCTs, we build an information theo
retic counterfactual variational information bottleneck (CVIB), as an alternativ
e for debiasing learning without RCTs. By separating the task-aware mutual infor
mation term in the original information bottleneck Lagrangian into factual and c
ounterfactual parts, we derive a contrastive information loss and an additional
output confidence penalty, which facilitates balanced learning between the factu
al and counterfactual domains. Empirical evaluation on real-world datasets shows
 that our CVIB significantly enhances both shallow and deep models, which sheds
light on counterfactual learning in recommendation that goes beyond RCTs.
**********
Prophet Attention: Predicting Attention with Future Attention
Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, Xu Sun
Recently, attention based models have been used extensively in many sequence-to-
sequence learning systems. Especially for image captioning, the attention based
models are expected to ground correct image regions with proper generated words.
 However, for each time step in the decoding process, the attention based models
 usually use the hidden state of the current input to attend to the image region
s. Under this setting, these attention models have a deviated focus'' problem th
at they calculate the attention weights based on previous words instead of the o
ne to be generated, impairing the performance of both grounding and captioning.
In this paper, we propose the Prophet Attention, similar to the form of self-sup
ervision. In the training stage, this module utilizes the future information to
calculate theideal'' attention weights towards image regions. These calculated i
deal'' weights are further used to regularize thedeviated'' attention. In this m
anner, image regions are grounded with the correct words. The proposed Prophet A
ttention can be easily incorporated into existing image captioning models to imp
rove their performance of both grounding and captioning. The experiments on the
Flickr30k Entities and the MSCOCO datasets show that the proposed Prophet Attent
ion consistently outperforms baselines in both automatic metrics and human evalu
ations. It is worth noticing that we set new state-of-the-arts on the two benchm
ark datasets and achieve the 1st place on the leaderboard of the online MSCOCO b
enchmark in terms of the default ranking score, i.e., CIDEr-c40.
**********
Language Models are Few-Shot Learners
Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla
Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhi
ni Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Adi
tya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen,
Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher
 Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei
We demonstrate that scaling up language models greatly improves task-agnostic, f
ew-shot performance, sometimes even becoming competitive with prior state-of-the
-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive lan
guage model with 175 billion parameters, 10x more than any previous non-sparse l
anguage model, and test its performance in the few-shot setting.  For all tasks,
 GPT-3 is applied without any gradient updates or fine-tuning, with tasks and fe
w-shot demonstrations specified purely via text interaction with the model.  GPT
-3 achieves strong performance on many NLP datasets, including translation, ques
tion-answering, and cloze tasks. We also identify some datasets where GPT-3's fe
w-shot learning still struggles, as well as some datasets where GPT-3 faces meth
odological issues related to training on large web corpora.
**********

Margins are Insufficient for Explaining Gradient Boosting
Allan Grønlund, Lior Kamma, Kasper Green Larsen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics
Alex Tseng, Avanti Shrikumar, Anshul Kundaje
Deep learning models can accurately map genomic DNA sequences to associated functional molecular readouts such as protein-DNA binding data. Base-resolution importance (i.e. "attribution") scores inferred from these models can highlight predictive sequence motifs and syntax. Unfortunately, these models are prone to overfitting and are sensitive to random initializations, often resulting in noisy and irreproducible attributions that obfuscate underlying motifs. To address these shortcomings, we propose a novel attribution prior, where the Fourier transform of input-level attribution scores are computed at training-time, and high-frequency components of the Fourier spectrum are penalized. We evaluate different model architectures with and without our attribution prior, training on genome-wide binary labels or continuous molecular profiles. We show that our attribution prior significantly improves models' stability, interpretability, and performance on held-out data, especially when training data is severely limited. Our attribution prior also allows models to identify biologically meaningful sequence motifs more sensitively and precisely within individual regulatory elements. The prior is agnostic to the model architecture or predicted experimental assay, yet provides similar gains across all experiments. This work represents an important advancement in improving the reliability of deep learning models for deciphering the regulatory code of the genome.
**********
MomentumRNN: Integrating Momentum into Recurrent Neural Networks
Tan Nguyen, Richard Baraniuk, Andrea Bertozzi, Stanley Osher, Bao Wang
Designing deep neural networks is an art that often involves an expensive search over candidate architectures. To overcome this for recurrent neural nets (RNNs), we establish a connection between the hidden state dynamics in an RNN and gradient descent (GD). We then integrate momentum into this framework and propose a new family of RNNs, called {\em MomentumRNNs}. We theoretically prove and numerically demonstrate that MomentumRNNs alleviate the vanishing gradient issue in training RNNs. We study the momentum long-short term memory (MomentumLSTM) and verify its advantages in convergence speed and accuracy over its LSTM counterpart across a variety of benchmarks. We also demonstrate that MomentumRNN is applicable to many types of recurrent cells, including those in the state-of-the-art orthogonal RNNs. Finally, we show that other advanced momentum-based optimization methods, such as Adam and Nesterov accelerated gradients with a restart, can be easily incorporated into the MomentumRNN framework for designing new recurrent cells with even better performance.
**********
Marginal Utility for Planning in Continuous or Large Discrete Action Spaces
Zaheen Ahmad, Levi Lelis, Michael Bowling
Sample-based planning is a powerful family of algorithms for generating intelligent behavior from a model of the environment. Generating good candidate actions is critical to the success of sample-based planners, particularly in continuous or large action spaces. Typically, candidate action generation exhausts the action space, uses domain knowledge, or more recently, involves learning a stochastic policy to provide such search guidance. In this paper we explore explicitly learning a candidate action generator by optimizing a novel objective, marginal utility. The marginal utility of an action generator measures the increase in value of an action over previously generated actions. We validate our approach in both curling, a challenging stochastic domain with continuous state and action spaces, and a location game with a discrete but large action space. We show that a

generator trained with the marginal utility objective outperforms hand-coded sch
emes built on substantial domain knowledge, trained stochastic policies, and oth
er natural objectives for generating actions for sampled-based planners.
**********
Projected Stein Variational Gradient Descent
Peng Chen, Omar Ghattas
The curse of dimensionality is a longstanding challenge in Bayesian
inference in high dimensions. In this work, we propose a {projected
  Stein variational gradient descent} (pSVGD) method to overcome this
challenge by exploiting the fundamental property of intrinsic low
dimensionality of the data informed subspace stemming from
ill-posedness of such problems. We adaptively construct the subspace
using a gradient information matrix of the log-likelihood, and apply
pSVGD to the much lower-dimensional coefficients of the parameter
projection. The method is demonstrated to be more accurate and
efficient than SVGD. It is also shown to be more scalable with respect
to the number of parameters, samples, data points, and processor cores
via experiments with parameters dimensions ranging from the
hundreds to the tens of thousands.
**********
Minimax Lower Bounds for Transfer Learning with Linear and One-hidden Layer Neur
al Networks
Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, Mahdi Soltanolkotab
i
Transfer learning has emerged as a powerful technique for improving the performa
nce of machine learning models on new domains where labeled training data may be
 scarce. In this approach a model trained for a source task, where plenty of lab
eled training data is available, is used as a starting point for training a mode
l on a related target task with only few labeled training data. Despite recent e
mpirical success of transfer learning approaches, the benefits and fundamental l
imits of transfer learning are poorly understood. In this paper we develop a sta
tistical minimax framework to characterize the fundamental limits of transfer le
arning in the context of regression with linear and one-hidden layer neural netw
ork models. Specifically, we derive a lower-bound for the target generalization
error achievable by any algorithm as a function of the number of labeled source
and target data as well as appropriate notions of similarity between the source
and target tasks. Our lower bound provides new insights into the benefits and li
mitations of transfer learning. We further corroborate our theoretical finding w
ith various experiments.
**********
SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks
Fabian Fuchs, Daniel Worrall, Volker Fischer, Max Welling
We introduce the SE(3)-Transformer, a variant of the self-attention module for 3
D point-clouds, which is equivariant under continuous 3D roto-translations. Equi
variance is important to ensure stable and predictable performance in the presen
ce of nuisance transformations of the data input. A positive corollary of equiva
riance is increased weight-tying within the model. The SE(3)-Transformer leverag
es the benefits of self-attention to operate on large point clouds with varying
number of points, while guaranteeing SE(3)-equivariance for robustness. We evalu
ate our model on a toy N-body particle simulation dataset, showcasing the robust
ness of the predictions under rotations of the input. We further achieve competi
tive performance on two real-world datasets, ScanObjectNN and QM9. In all cases,
 our model outperforms a strong, non-equivariant attention baseline and an equiv
ariant model without attention.
**********
On the equivalence of molecular graph convolution and molecular wave function wi
th poor basis set
Masashi Tsubaki, Teruyasu Mizoguchi
In this study, we demonstrate that the linear combination of atomic orbitals (LC
AO), an approximation introduced by Pauling and Lennard-Jones in the 1920s, corr

esponds to graph convolutional networks (GCNs) for molecules. However, GCNs involve unnecessary nonlinearity and deep architecture. We also verify that molecular GCNs are based on a poor basis function set compared with the standard one used in theoretical calculations or quantum chemical simulations. From these observations, we describe the quantum deep field (QDF), a machine learning (ML) model based on an underlying quantum physics, in particular the density functional theory (DFT). We believe that the QDF model can be easily understood because it can be regarded as a single linear layer GCN. Moreover, it uses two vanilla feedforward neural networks to learn an energy functional and a Hohenberg--Kohn map that have nonlinearities inherent in quantum physics and the DFT. For molecular energy prediction tasks, we demonstrated the viability of an ``extrapolation,'' in which we trained a QDF model with small molecules, tested it with large molecules, and achieved high extrapolation performance. We believe that we should move away from the competition of interpolation accuracy within benchmark datasets and evaluate ML models based on physics using an extrapolation setting; this will lead to reliable and practical applications, such as fast, large-scale molecular screening for discovering effective materials.
**********

The Power of Predictions in Online Control
Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, 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-authors prior to requesting a name change in the electronic proceedings.
**********

Learning Affordance Landscapes for Interaction Exploration in 3D Environments
Tushar Nagarajan, Kristen Grauman
Embodied agents operating in human spaces must be able to master how their environment works: what objects can the agent use, and how can it use them? We introduce a reinforcement learning approach for exploration for interaction, whereby an embodied agent autonomously discovers the affordance landscape of a new unmapped 3D environment (such as an unfamiliar kitchen). Given an egocentric RGB-D camera and a high-level action space, the agent is rewarded for maximizing successful interactions while simultaneously training an image-based affordance segmentation model. The former yields a policy for acting efficiently in new environments to prepare for downstream interaction tasks, while the latter yields a convolutional neural network that maps image regions to the likelihood they permit each action, densifying the rewards for exploration. We demonstrate our idea with AI2-iTHOR. The results show agents can learn how to use new home environments intelligently and that it prepares them to rapidly address various downstream tasks like "find a knife and put it in the drawer." Project page: http://vision.cs.utexas.edu/projects/interaction-exploration/
**********

Cooperative Multi-player Bandit Optimization
Ilai Bistritz, Nicholas Bambos
Consider a team of cooperative players that take actions in a networked-environment. At each turn, each player chooses an action and receives a reward that is an unknown function of all the players' actions. The goal of the team of players is to learn to play together the action profile that maximizes the sum of their rewards. However, players cannot observe the actions or rewards of other players, and can only get this information by communicating with their neighbors. We design a distributed learning algorithm that overcomes the informational bias players have towards maximizing the rewards of nearby players they got more information about. We assume twice continuously differentiable reward functions and constrained convex and compact action sets. Our communication graph is a random time-varying graph that follows an ergodic Markov chain. We prove that even if at every turn players take actions based only on the small random subset of the players' rewards that they know, our algorithm converges with probability 1 to the set of stationary points of (projected) gradient ascent on the sum of rewards function. Hence, if the sum of rewards is concave, then the algorithm converges with

probability 1 to the optimal action profile.
**********
Tight First- and Second-Order Regret Bounds for Adversarial Linear Bandits
Shinji Ito, Shuichi Hirahara, Tasuku Soma, Yuichi Yoshida
We propose novel algorithms with first- and second-order regret bounds for adversarial linear bandits. These regret bounds imply that our algorithms perform well when there is an action achieving a small cumulative loss or the loss has a small variance. In addition, we need only assumptions weaker than those of existing algorithms; our algorithms work on discrete action sets as well as continuous ones without a priori knowledge about losses, and they run efficiently if a linear optimization oracle for the action set is available. These results are obtained by combining optimistic online optimization, continuous multiplicative weight update methods, and a novel technique that we refer to as distribution truncation. We also show that the regret bounds of our algorithms are tight up to polylogarithmic factors.
**********
Just Pick a Sign: Optimizing Deep Multitask Models with Gradient Sign Dropout
Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, Dragomir Anguelov
The vast majority of deep models use multiple gradient signals, typically corresponding to a sum of multiple loss terms, to update a shared set of trainable weights. However, these multiple updates can impede optimal training by pulling the model in conflicting directions. We present Gradient Sign Dropout (GradDrop), a probabilistic masking procedure which samples gradients at an activation layer based on their level of consistency. GradDrop is implemented as a simple deep layer that can be used in any deep net and synergizes with other gradient balancing approaches. We show that GradDrop outperforms the state-of-the-art multiloss methods within traditional multitask and transfer learning settings, and we discuss how GradDrop reveals links between optimal multiloss training and gradient stochasticity.
**********
A Loss Function for Generative Neural Networks Based on Watson’s Perceptual Model
Steffen Czolbe, Oswin Krause, Ingemar Cox, Christian Igel
To train  Variational Autoencoders (VAEs) to generate realistic imagery requires a loss function that reflects human perception of image similarity. We propose such a loss function based on Watson's perceptual model, which computes a weighted distance in frequency space and accounts for luminance and contrast masking. We extend the model to color images, increase its robustness to translation by using the Fourier Transform, remove artifacts due to splitting the image into blocks, and make it differentiable.
In experiments, VAEs trained with the new loss function generated realistic, high-quality image samples. Compared to using the Euclidean distance and the Structural Similarity Index, the images were less blurry; compared to deep neural network based losses, the new approach required less computational resources and generated images with less artifacts.
**********
Dynamic Fusion of Eye Movement Data and Verbal Narrations in Knowledge-rich Domains
Ervine Zheng, Qi Yu, Rui Li, Pengcheng Shi, Anne Haake
We propose to jointly analyze experts' eye movements and verbal narrations to discover important and interpretable knowledge patterns to better understand their decision-making processes. The discovered patterns can further enhance data-driven statistical models by fusing experts' domain knowledge to support complex human-machine collaborative decision-making. Our key contribution is a novel dynamic Bayesian nonparametric model that assigns latent knowledge patterns into key phases involved in complex decision-making. Each phase is characterized by a unique distribution of word topics discovered from verbal narrations and their dynamic interactions with eye movement patterns, indicating experts' special perceptual behavior within a given decision-making stage. A new split-merge-switch samp

ler is developed to efficiently explore the posterior state space with an improv ed mixing rate. Case studies on diagnostic error prediction and disease morpholo gy categorization help demonstrate the effectiveness of the proposed model and d iscovered knowledge patterns.
**********

Scalable Multi-Agent Reinforcement Learning for Networked Systems with Average R eward
Guannan Qu, Yiheng Lin, Adam Wierman, Na Li
It has long been recognized that multi-agent reinforcement learning (MARL) faces  significant scalability issues due to the fact that the size of the state and a ction spaces are exponentially large in the number of agents. In this paper, we identify a rich class of networked MARL problems where the model exhibits a loca l dependence structure that allows it to be solved in a scalable manner. Specifi cally, we propose a Scalable Actor-Critic (SAC) method that can learn a near opt imal localized policy for optimizing the average reward with complexity scaling with the state-action space size of local neighborhoods, as opposed to the entir e network. Our result centers around identifying and exploiting an exponential d ecay property that ensures the effect of agents on each other decays exponential ly fast in their graph distance.
**********

Optimizing Neural Networks via Koopman Operator Theory
Akshunna S. Dogra, William Redman
Koopman operator theory, a powerful framework for discovering the underlying dyn amics of nonlinear dynamical systems, was recently shown to be intimately connec ted with neural network training. In this work, we take the first steps in makin g use of this connection. As Koopman operator theory is a linear theory, a succe ssful implementation of it in evolving network weights and biases offers the pro mise of accelerated training, especially in the context of deep networks, where optimization is inherently a non-convex problem. We show that Koopman operator t heoretic methods allow for accurate predictions of weights and biases of feedfor ward, fully connected deep networks over a non-trivial range of training time. D uring this window, we find that our approach is >10x faster than various gradien t descent based methods (e.g. Adam, Adadelta, Adagrad), in line with our complex ity analysis. We end by highlighting open questions in this exciting intersectio n between dynamical systems and neural network theory. We highlight additional m ethods by which our results could be expanded to broader classes of networks and  larger training intervals, which shall be the focus of future work.
**********

SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence
Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet
Stein Variational Gradient Descent (SVGD), a popular sampling algorithm, is ofte n described as the kernelized gradient flow for the Kullback-Leibler divergence in the geometry of optimal transport. We introduce a new perspective on SVGD tha t instead views SVGD as the kernelized gradient flow of the chi-squared divergen ce. Motivated by this perspective, we provide a convergence analysis of the chi-squared gradient flow. We also show that our new perspective provides better gui delines for choosing effective kernels for SVGD.
**********

Adversarial Robustness of Supervised Sparse Coding
Jeremias Sulam, Ramchandran Muthukumar, 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-auth ors prior to requesting a name change in the electronic proceedings.
**********

Differentiable Meta-Learning of Bandit Policies
Craig Boutilier, Chih-wei Hsu, Branislav Kveton, Martin Mladenov, Csaba Szepesva ri, Manzil Zaheer
Exploration policies in Bayesian bandits maximize the average reward over proble m instances drawn from some distribution P. In this work, we learn such policies

for an unknown distribution P using samples from P. Our approach is a form of m
eta-learning and exploits properties of P without making strong assumptions abou
t its form. To do this, we parameterize our policies in a differentiable way and
optimize them by policy gradients, an approach that is pleasantly general and e
asy to implement. We derive effective gradient estimators and propose novel vari
ance reduction techniques. We also analyze and experiment with various bandit po
licy classes, including neural networks and a novel softmax policy. The latter h
as regret guarantees and is a natural starting point for our optimization. Our e
xperiments show the versatility of our approach. We also observe that neural net
work policies can learn implicit biases expressed only through the sampled insta
nces.
**********

Biologically Inspired Mechanisms for Adversarial Robustness
Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, Tomaso Poggio
A convolutional neural network strongly robust to adversarial perturbations at r
easonable computational and performance cost has not yet been demonstrated. The
primate visual ventral stream seems to be robust to small perturbations in visua
l stimuli but the underlying mechanisms that give rise to this robust perception
are not understood. In this work, we investigate the role of two biologically p
lausible mechanisms in adversarial robustness. We demonstrate that the non-unifo
rm sampling performed by the primate retina and the presence of multiple recepti
ve fields with a range of receptive field sizes at each eccentricity improve the
robustness of neural networks to small adversarial perturbations. We verify tha
t these two mechanisms do not suffer from gradient obfuscation and study their c
ontribution to adversarial robustness through ablation studies.
**********

Statistical-Query Lower Bounds via Functional Gradients
Surbhi Goel, Aravind Gollakota, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Near-Optimal Reinforcement Learning with Self-Play
Yu Bai, Chi Jin, Tiancheng Yu
This paper considers the problem of designing optimal algorithms for reinforceme
nt learning in two-player zero-sum games. We focus on self-play algorithms which
learn the optimal policy by playing against itself without any direct supervisi
on. In a tabular episodic Markov game with S states, A max-player actions and B
min-player actions, the best existing algorithm for finding an approximate Nash
equilibrium requires \tlO(S^2AB) steps of game playing, when only highlighting t
he dependency on (S,A,B). In contrast, the best existing lower bound scales as \
Omega(S(A+B)) and has a significant gap from the upper bound. This paper closes
this gap for the first time: we propose an optimistic variant of the Nash Q-lear
ning algorithm with sample complexity \tlO(SAB), and a new Nash V-learning algor
ithm with sample complexity \tlO(S(A+B)). The latter result matches the informat
ion-theoretic lower bound in all problem-dependent parameters except for a polyn
omial factor of the length of each episode. In addition, we present a computatio
nal hardness result for learning the best responses against a fixed opponent in
Markov games---a learning objective different from finding the Nash equilibrium.
**********

Network Diffusions via Neural Mean-Field Dynamics
Shushan He, Hongyuan Zha, Xiaojing Ye
We propose a novel learning framework based on neural mean-field dynamics for in
ference and estimation problems of diffusion on networks. Our new framework is d
erived from the Mori-Zwanzig formalism to obtain an exact evolution of the node
infection probabilities, which renders a delay differential equation with memory
integral approximated by learnable time convolution operators, resulting in a h
ighly structured and interpretable RNN. Directly using cascade data, our framewo
rk can jointly learn the structure of the diffusion network and the evolution of

infection probabilities, which are cornerstone to important downstream applications such as influence maximization. Connections between parameter learning and optimal control are also established. Empirical study shows that our approach is versatile and robust to variations of the underlying diffusion network models, and significantly outperform existing approaches in accuracy and efficiency on both synthetic and real-world data.

**********

Self-Distillation as Instance-Specific Label Smoothing

Zhilu Zhang, Mert Sabuncu

It has been recently demonstrated that multi-generational self-distillation can improve generalization. Despite this intriguing observation, reasons for the enhancement remain poorly understood. In this paper, we first demonstrate experimentally that the improved performance of multi-generational self-distillation is in part associated with the increasing diversity in teacher predictions. With this in mind, we offer a new interpretation for teacher-student training as amortized MAP estimation, such that teacher predictions enable instance-specific regularization. Our framework allows us to theoretically relate self-distillation to label smoothing, a commonly used technique that regularizes predictive uncertainty, and suggests the importance of predictive diversity in addition to predictive uncertainty. We present experimental results using multiple datasets and neural network architectures that, overall, demonstrate the utility of predictive diversity. Finally, we propose a novel instance-specific label smoothing technique that promotes predictive diversity without the need for a separately trained teacher model. We provide an empirical evaluation of the proposed method, which, we find, often outperforms classical label smoothing.

**********

Towards Problem-dependent Optimal Learning Rates

Yunbei Xu, Assaf Zeevi

We study problem-dependent rates, i.e., generalization errors that scale tightly with the variance or the effective loss at the "best hypothesis." Existing uniform convergence and localization frameworks, the most widely used tools to study this problem, often fail to simultaneously provide parameter localization and optimal dependence on the sample size. As a result, existing problem-dependent rates are often rather weak when the hypothesis class is "rich" and the worst-case bound of the loss is large. In this paper we propose a new framework based on a "uniform localized convergence" principle. We provide the first (moment-penalized) estimator that achieves the optimal variance-dependent rate for general "rich" classes; we also establish improved loss-dependent rate for standard empirical risk minimization.

**********

Cross-lingual Retrieval for Iterative Self-Supervised Training

Chau Tran, Yuqing Tang, Xian Li, Jiatao Gu

Recent studies have demonstrated the cross-lingual alignment ability of multilingual pretrained language models. In this work, we found that the cross-lingual alignment can be further improved by training seq2seq models on sentence pairs mined using their own encoder outputs. We utilized these findings to develop a new approach --- cross-lingual retrieval for iterative self-supervised training (CRISS), where mining and training processes are applied iteratively, improving cross-lingual alignment and translation ability at the same time. Using this method, we achieved state-of-the-art unsupervised machine translation results on 9 language directions with an average improvement of 2.4 BLEU, and on the Tatoeba sentence retrieval task in the XTREME benchmark on 16 languages with an average improvement of 21.5% in absolute accuracy. Furthermore, CRISS also brings an additional 1.8 BLEU improvement on average compared to mBART, when finetuned on supervised machine translation downstream tasks.

**********

Rethinking pooling in graph neural networks

Diego Mesquita, Amauri Souza, Samuel Kaski

Graph pooling is a central component of a myriad of graph neural network (GNN) architectures. As an inheritance from traditional CNNs, most approaches formulate

graph pooling as a cluster assignment problem, extending the idea of local patc
hes in regular grids to graphs. Despite the wide adherence to this design choice
, no work has rigorously evaluated its influence on the success of GNNs. In this
 paper, we build upon representative GNNs and introduce variants that challenge
the need for locality-preserving representations, either using randomization or
clustering on the complement graph. Strikingly, our experiments demonstrate that
 using these variants does not result in any decrease in performance. To underst
and this phenomenon, we study the interplay between convolutional layers and the
 subsequent pooling ones. We show that the convolutions play a leading role in t
he learned representations. In contrast to the common belief, local pooling is n
ot responsible for the success of GNNs on relevant and widely-used benchmarks.
**********

Pointer Graph Networks

Petar Veli■kovi■, Lars Buesing, Matthew Overlan, Razvan Pascanu, Oriol Vinyals,
Charles Blundell

Graph neural networks (GNNs) are typically applied to static graphs that are ass
umed to be known upfront. This static input structure is often informed purely b
y insight of the machine learning practitioner, and might not be optimal for the
 actual task the GNN is solving. In absence of reliable domain expertise, one mi
ght resort to inferring the latent graph structure, which is often difficult due
 to the vast search space of possible graphs. Here we introduce Pointer Graph Ne
tworks (PGNs) which augment sets or graphs with additional inferred edges for im
proved model generalisation ability. PGNs allow each node to dynamically point t
o another node, followed by message passing over these pointers. The sparsity of
 this adaptable graph structure makes learning tractable while still being suffi
ciently expressive to simulate complex algorithms. Critically, the pointing mech
anism is directly supervised to model long-term sequences of operations on class
ical data structures, incorporating useful structural inductive biases from theo
retical computer science. Qualitatively, we demonstrate that PGNs can learn para
llelisable variants of pointer-based data structures, namely disjoint set unions
 and link/cut trees. PGNs generalise out-of-distribution to 5x larger test input
s on dynamic graph connectivity tasks, outperforming unrestricted GNNs and Deep
Sets.
**********

Gradient Regularized V-Learning for Dynamic Treatment Regimes

Yao Zhang, Mihaela van der Schaar

Deciding how to optimally treat a patient, including how to select treatments ov
er time among the multiple available treatments, represents one of the most impo
rtant issues that need to be addressed in medicine today. A dynamic treatment re
gime (DTR) is a sequence of treatment rules indicating how to individualize trea
tments for a patient based on the previously assigned treatments and the evolvin
g covariate history. However, DTR evaluation and learning based on offline data
remain challenging problems due to the bias introduced by time-varying confounde
rs that affect treatment assignment over time; this may lead to suboptimal treat
ment rules being used in practice. In this paper, we introduce Gradient Regulari
zed V-learning (GRV), a novel method for estimating the value function of a DTR.
 GRV regularizes the underlying outcome and propensity score models with respect
 to the optimality condition in semiparametric estimation theory. On the basis o
f this design, we construct estimators that are efficient and stable in finite s
amples regime. Using multiple simulation studies and one real-world medical data
set, we demonstrate that our method is superior in DTR evaluation and learning,
thereby providing improved treatment options over time for patients.
**********

Faster Wasserstein Distance Estimation with the Sinkhorn Divergence

Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, Gabrie
l Peyré

Requests for name changes in the electronic proceedings will be accepted with 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.

**********
Forethought and Hindsight in Credit Assignment
Veronica Chelu, Doina Precup, Hado P. van Hasselt
We address the problem of credit assignment in reinforcement learning and explore fundamental questions regarding the way in which an agent can best use additional computation to propagate new information, by planning with internal models of the world to improve its predictions. Particularly, we work to understand the gains and peculiarities of planning employed as forethought via forward models or as hindsight operating with backward models. We establish the relative merits, limitations and complementary properties of both planning mechanisms in carefully constructed scenarios. Further, we investigate the best use of models in planning, primarily focusing on the selection of states in which predictions should be (re)-evaluated. Lastly, we discuss the issue of model estimation and highlight a spectrum of methods that stretch from environment dynamics predictors to planner-aware models.
**********
Robust Recursive Partitioning for Heterogeneous Treatment Effects with Uncertainty Quantification
Hyun-Suk Lee, Yao Zhang, William Zame, Cong Shen, Jang-Won Lee, Mihaela van der Schaar
Subgroup analysis of treatment effects plays an important role in applications from medicine to public policy to recommender systems. It allows physicians (for example) to identify groups of patients for whom a given drug or treatment is likely to be effective and groups of patients for which it is not. Most of the current methods of subgroup analysis begin with a particular algorithm for estimating individualized treatment effects (ITE) and identify subgroups by maximizing the difference across subgroups of the average treatment effect in each subgroup. These approaches have several weaknesses: they rely on a particular algorithm for estimating ITE, they ignore (in)homogeneity within identified subgroups, and they do not produce good confidence estimates. This paper develops a new method for subgroup analysis, R2P, that addresses all these weaknesses. R2P uses an arbitrary, exogenously prescribed algorithm for estimating ITE and quantifies the uncertainty of the ITE estimation, using a construction that is more robust than other methods. Experiments using synthetic and semi-synthetic datasets (based on real data) demonstrate that R2P constructs partitions that are simultaneously more homogeneous within groups and more heterogeneous across groups than the partitions produced by other methods. Moreover, because R2P can employ any ITE estimator, it also produces much narrower confidence intervals with a prescribed coverage guarantee than other methods.
**********
Rescuing neural spike train models from bad MLE
Diego Arribas, Yuan Zhao, Il Memming Park
The standard approach to fitting an autoregressive spike train model is to maximize the likelihood for one-step prediction. This maximum likelihood estimation (MLE) often leads to models that perform poorly when generating samples recursively for more than one time step. Moreover, the generated spike trains can fail to capture important features of the data and even show diverging firing rates. To alleviate this, we propose to directly minimize the divergence between neural recorded and model generated spike trains using spike train kernels. We develop a method that stochastically optimizes the maximum mean discrepancy induced by the kernel. Experiments performed on both real and synthetic neural data validate the proposed approach, showing that it leads to well-behaving models. Using different combinations of spike train kernels, we show that we can control the trade-off between different features which is critical for dealing with model-mismatch.
**********
Lower Bounds and Optimal Algorithms for Personalized Federated Learning
Filip Hanzely, Slavomír Hanzely, Samuel Horváth, Peter Richtarik
In this work, we consider the optimization formulation of personalized federated learning recently introduced by Hanzely & Richtarik (2020) which was shown to g

ive an alternative explanation to the workings of local SGD methods. Our first c
ontribution is establishing the first lower bounds for this formulation, for bot
h the communication complexity and the local oracle complexity. Our second contr
ibution is the design of several optimal methods matching these lower bounds in
almost all regimes. These are the first provably optimal methods for personalize
d federated learning. Our optimal methods include an accelerated variant of FedP
rox, and an accelerated variance-reduced version of FedAvg/Local SGD. We demonst
rate the practical superiority of our methods through extensive numerical experi
ments.
**********

Black-Box Certification with Randomized Smoothing: A Functional Optimization Bas
ed Framework
Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Deep Imitation Learning for Bimanual Robotic Manipulation
Fan Xie, Alexander Chowdhury, M. Clara De Paolis Kaluza, Linfeng Zhao, Lawson Wo
ng, Rose Yu
We present a deep imitation learning framework for robotic bimanual manipulation
 in a continuous state-action space.  A core challenge is to generalize the mani
pulation skills to objects in different locations. We hypothesize that modeling
the relational information in the environment can significantly improve generali
zation. To achieve this, we propose to (i) decompose the multi-modal dynamics in
to elemental movement primitives, (ii) parameterize each primitive using a recur
rent graph neural network to capture interactions, and (iii) integrate a high-le
vel planner that composes primitives sequentially and a low-level controller to
combine primitive dynamics and inverse kinematics control.  Our model is a deep,
 hierarchical, modular architecture. Compared to baselines, our model generalize
s better and achieves higher success rates on several simulated bimanual robotic
 manipulation tasks. We open source the code for simulation, data, and models at
: https://github.com/Rose-STL-Lab/HDR-IL.
**********

Stationary Activations for Uncertainty Calibration in Deep Learning
Lassi Meronen, Christabella Irwanto, Arno Solin
We introduce a new family of non-linear neural network activation functions that
 mimic the properties induced by the widely-used Mat\'ern family of kernels in G
aussian process (GP) models. This class spans a range of locally stationary mode
ls of various degrees of mean-square differentiability.  We show an explicit lin
k to the corresponding GP models in the case that the network consists of one in
finitely wide hidden layer. In the limit of infinite smoothness the Mat\'ern fam
ily results in the RBF kernel, and in this case we recover RBF activations. Mat\
'ern activation functions result in similar appealing properties to their counte
rparts in GP models, and we demonstrate that the local stationarity property tog
ether with limited mean-square differentiability shows both good performance and
 uncertainty calibration in Bayesian deep learning tasks. In particular, local s
tationarity helps calibrate out-of-distribution (OOD) uncertainty. We demonstrat
e these properties on classification and regression benchmarks and a radar emitt
er classification task.
**********

Ensemble Distillation for Robust Model Fusion in Federated Learning
Tao Lin, Lingjing Kong, Sebastian U. Stich, Martin Jaggi
Federated Learning (FL) is a machine learning setting where many devices collabo
ratively train a machine learning model while keeping the training data decentra
lized. In most of the current training schemes the central model is refined by a
veraging the parameters of the server model and the updated parameters from the
client side. However, directly averaging model parameters is only possible if al
l models have the same structure and size, which could be a restrictive constrai

nt in many scenarios.
**********
Falcon: Fast Spectral Inference on Encrypted Data
Qian Lou, Wen-jie Lu, Cheng Hong, Lei Jiang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Power Laws in Deep Ensembles
Ekaterina Lobacheva, Nadezhda Chirkova, Maxim Kodryan, Dmitry P. Vetrov
Ensembles of deep neural networks are known to achieve state-of-the-art performance in uncertainty estimation and lead to accuracy improvement. In this work, we focus on a classification problem and investigate the behavior of both non-calibrated and calibrated negative log-likelihood (CNLL) of a deep ensemble as a function of the ensemble size and the member network size. We indicate the conditions under which CNLL follows a power law w. r. t. ensemble size or member network size, and analyze the dynamics of the parameters of the discovered power laws. Our important practical finding is that one large network may perform worse than an ensemble of several medium-size networks with the same total number of parameters (we call this ensemble a memory split). Using the detected power law-like dependencies, we can predict (1) the possible gain from the ensembling of networks with given structure, (2) the optimal memory split given a memory budget, based on a relatively small number of trained networks.
**********
Practical Quasi-Newton Methods for Training Deep Neural Networks
Donald Goldfarb, Yi Ren, Achraf Bahamou
We consider the development of practical stochastic quasi-Newton, and in particular Kronecker-factored block diagonal BFGS and L-BFGS methods, for training deep neural networks (DNNs). In DNN training, the number of variables and components of the gradient n is often of the order of tens of millions and the Hessian has n^2 elements. Consequently, computing and storing a full n times n BFGS approximation or storing a modest number of (step, change in gradient) vector pairs for use in an L-BFGS implementation is out of the question. In our proposed methods, we approximate the Hessian by a block-diagonal matrix and use the structure of the gradient and Hessian to further approximate these blocks, each of which corresponds to a layer, as the Kronecker product of two much smaller matrices. This is analogous to the approach in KFAC , which computes a Kronecker-factored block diagonal approximation to the Fisher matrix in a stochastic natural gradient method. Because the indefinite and highly variable nature of the Hessian in a DNN, we also propose a new damping approach to keep the upper as well as the lower bounds of the BFGS and L-BFGS approximations bounded. In tests on autoencoder feed-forward network models with either nine or thirteen layers applied to three datasets, our methods outperformed or performed comparably to KFAC and state-of-the-art first-order stochastic methods.
**********
Approximation Based Variance Reduction for Reparameterization Gradients
Tomas Geffner, Justin Domke
Flexible variational distributions improve variational inference but are harder to optimize. In this work we present a control variate that is applicable for any reparameterizable distribution with known mean and covariance, e.g. Gaussians with any covariance structure. The control variate is based on a quadratic approximation of the model, and its parameters are set using a double-descent scheme. We empirically show that this control variate leads to large improvements in gradient variance and optimization convergence for inference with non-factorized variational distributions.
**********
Inference Stage Optimization for Cross-scenario 3D Human Pose Estimation
Jianfeng Zhang, Xuecheng Nie, Jiashi Feng
Existing 3D human pose estimation models suffer performance drop when applying t

o new scenarios with unseen poses due to their limited generalizability. In this work, we propose a novel framework, Inference Stage Optimization (ISO), for improving the generalizability of 3D pose models when source and target data come from different pose distributions. Our main insight is that the target data, even though not labeled, carry valuable priors about their underlying distribution. To exploit such information, the proposed ISO performs geometry-aware self-supervised learning (SSL) on each single target instance and updates the 3D pose model before making prediction. In this way, the model can mine distributional knowledge about the target scenario and quickly adapt to it with enhanced generalization performance. In addition, to handle sequential target data, we propose an online mode for implementing our ISO framework via streaming the SSL, which substantially enhances its effectiveness. We systematically analyze why and how our ISO framework works on diverse benchmarks under cross-scenario setup. Remarkably, it yields new state-of-the-art of 83.6% 3D PCK on MPI-INF-3DHP, improving upon the previous best result by 9.7%.

**********

Consistent feature selection for analytic deep neural networks

Vu C. Dinh, Lam S. Ho

One of the most important steps toward interpretability and explainability of neural network models is feature selection, which aims to identify the subset of relevant features. Theoretical results in the field have mostly focused on the prediction aspect of the problem with virtually no work on feature selection consistency for deep neural networks due to the model's severe nonlinearity and unidentifiability. This lack of theoretical foundation casts doubt on the applicability of deep learning to contexts where correct interpretations of the features play a central role.

**********

Glance and Focus: a Dynamic Approach to Reducing Spatial Redundancy in Image Classification

Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, Gao Huang

The accuracy of deep convolutional neural networks (CNNs) generally improves when fueled with high resolution images. However, this often comes at a high computational cost and high memory footprint. Inspired by the fact that not all regions in an image are task-relevant, we propose a novel framework that performs efficient image classification by processing a sequence of relatively small inputs, which are strategically selected from the original image with reinforcement learning. Such a dynamic decision process naturally facilitates adaptive inference at test time, i.e., it can be terminated once the model is sufficiently confident about its prediction and thus avoids further redundant computation. Notably, our framework is general and flexible as it is compatible with most of the state-of-the-art light-weighted CNNs (such as MobileNets, EfficientNets and RegNets), which can be conveniently deployed as the backbone feature extractor. Experiments on ImageNet show that our method consistently improves the computational efficiency of a wide variety of deep models. For example, it further reduces the average latency of the highly efficient MobileNet-V3 on an iPhone XS Max by 20% without sacrificing accuracy. Code and pre-trained models are available at https://github.com/blackfeather-wang/GFNet-Pytorch.

**********

Information Maximization for Few-Shot Learning

Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, Jose Dolz, Pablo Piantanida, Ismail Ben Ayed

We introduce Transductive Infomation Maximization (TIM) for few-shot learning. Our method maximizes the mutual information between the query features and their label predictions for a given few-shot task, in conjunction with a supervision loss based on the support set. Furthermore, we propose a new alternating-direction solver for our mutual-information loss, which substantially speeds up transductive inference convergence over gradient-based optimization, while yielding similar accuracy.  TIM inference is modular:  it can be used on top of any base-training feature extractor. Following standard transductive few-shot settings, our comprehensive experiments demonstrate that TIM outperforms state-of-the-art metho

ds significantly across various datasets and networks, while used on top of a fi
xed feature extractor trained with simple cross-entropy on the base classes, wit
hout resorting to complex meta-learning schemes. It consistently brings between
2% and 5% improvement in accuracy over the best performing method, not only on a
ll the well-established few-shot benchmarks but also on more challenging scenari
os, with domain shifts and larger numbers of classes.
**********

Inverse Reinforcement Learning from a Gradient-based Learner
Giorgia Ramponi, Gianluca Drappo, Marcello Restelli
Inverse Reinforcement Learning addresses the problem of inferring an expert's re
ward function from demonstrations. However, in many applications, we not only ha
ve access to the expert's near-optimal behaviour, but we also observe part of he
r learning process.
In this paper, we propose a new algorithm for this setting, in which the goal is
 to recover the reward function being optimized by an agent, given a sequence of
 policies produced during learning. Our approach is based on the assumption that
 the observed agent is updating her policy parameters along the gradient directi
on. Then we extend our method to deal with the more realistic scenario where we
only have access to a dataset of learning trajectories. For both settings, we pr
ovide theoretical insights into our algorithms' performance. Finally, we evaluat
e the approach in a simulated GridWorld environment and on the MuJoCo environmen
ts, comparing it with the state-of-the-art baseline.
**********

Bayesian Multi-type Mean Field Multi-agent Imitation Learning
Fan Yang, Alina Vereshchaka, Changyou Chen, Wen Dong
Multi-agent Imitation learning (MAIL) refers to the problem that agents learn to
 perform a task interactively in a multi-agent system through observing and mimi
cking expert demonstrations, without any knowledge of a reward function from the
 environment. MAIL has received a lot of attention due to promising results achi
eved on synthesized tasks, with the potential to be applied to complex real-worl
d multi-agent tasks. Key challenges for MAIL include sample efficiency and scala
bility. In this paper, we proposed Bayesian multi-type mean field multi-agent im
itation learning (BM3IL). Our method improves sample efficiency through establis
hing a Bayesian formulation for MAIL, and enhances scalability through introduci
ng a new multi-type mean field approximation. We demonstrate the performance of
our algorithm through benchmarking with three state-of-the-art multi-agent imita
tion learning algorithms on several tasks, including solving a multi-agent traff
ic optimization problem in a real-world transportation network. Experimental res
ults indicate that our algorithm significantly outperforms all other algorithms
in all scenarios.
**********

Bayesian Robust Optimization for Imitation Learning
Daniel Brown, Scott Niekum, Marek Petrik
One of the main challenges in imitation learning is determining what action an a
gent should take when outside the state distribution of the demonstrations. Inve
rse reinforcement learning (IRL) can enable generalization to new states by lear
ning a parameterized reward function, but these approaches still face uncertaint
y over the true reward function and corresponding optimal policy.  Existing safe
 imitation learning approaches based on IRL deal with this uncertainty using a m
axmin framework that optimizes a policy under the assumption of an adversarial r
eward function, whereas risk-neutral IRL approaches either optimize a policy for
 the mean or MAP reward function.  While completely ignoring risk can lead to ov
erly aggressive and unsafe policies, optimizing in a fully adversarial sense is
also problematic as it can lead to overly conservative policies that perform poo
rly in practice. To provide a bridge between these two extremes,  we propose Bay
esian Robust Optimization for Imitation Learning (BROIL). BROIL leverages Bayesi
an reward function inference and a user specific risk tolerance to efficiently o
ptimize a robust policy that balances expected return and conditional value at r
isk. Our empirical results show that BROIL provides a natural way to interpolate
 between return-maximizing and risk-minimizing behaviors and outperforms existin

g risk-sensitive and risk-neutral inverse reinforcement learning algorithms.
**********
Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance
Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, Ya ron Lipman
In this work we address the challenging problem of multiview 3D surface reconstr uction. We introduce a neural network architecture that simultaneously learns th e unknown geometry, camera parameters, and a neural renderer that approximates t he light reflected from the surface towards the camera.
The geometry is represented as a zero level-set of a neural network, while the n eural renderer, derived from the rendering equation, is capable of (implicitly) modeling a wide set of lighting conditions and materials.
We trained our network on real world 2D images of objects with different materia l properties, lighting conditions, and noisy camera initializations from the DTU MVS dataset. We found our model to produce state of the art 3D surface reconstr uctions with high fidelity, resolution and detail.
**********
Riemannian Continuous Normalizing Flows
Emile Mathieu, Maximilian Nickel
Normalizing flows have shown great promise for modelling flexible probability di stributions in a computationally tractable way. However, whilst data is often na turally described on Riemannian manifolds such as spheres, torii, and hyperbolic spaces, most normalizing flows implicitly assume a flat geometry, making them e ither misspecified or ill-suited in these situations. To overcome this problem, we introduce Riemannian continuous normalizing flows, a model which admits the p arametrization of flexible probability measures on smooth manifolds by defining flows as the solution to ordinary differential equations. We show that this appr oach can lead to substantial improvements on both synthetic and real-world data when compared to standard flows or previously introduced projected flows.
**********
Attention-Gated Brain Propagation: How the brain can implement reward-based erro r backpropagation
Isabella Pozzi, Sander  Bohte, Pieter Roelfsema
Much recent work has focused on biologically plausible variants of supervised le arning algorithms. However, there is no teacher in the motor cortex that instruc ts the motor neurons and learning in the brain depends on reward and punishment. We demonstrate a biologically plausible reinforcement learning scheme for deep networks with an arbitrary number of layers. The network chooses an action by se lecting a unit in the output layer and uses feedback connections to assign credi t to the units in successively lower layers that are responsible for this action . After the choice, the network receives reinforcement and there is no teacher c orrecting the errors. We show how the new learning scheme – Attention-Gated Brai n Propagation (BrainProp) – is mathematically equivalent to error backpropagatio n, for one output unit at a time. We demonstrate successful learning of deep ful ly connected, convolutional and locally connected networks on classical and hard image-classification benchmarks; MNIST, CIFAR10, CIFAR100 and Tiny ImageNet. Br ainProp achieves an accuracy that is equivalent to that of standard error-backpr opagation, and better than state-of-the-art biologically inspired learning schem es. The trial-and-error nature of learning is associated with limited additional training time so that BrainProp is a factor of 1-3.5 times slower. Our results thereby provide new insights into how deep learning may be implemented in the br ain.
**********
Asymptotic Guarantees for Generative Modeling Based on the Smooth Wasserstein Di stance
Ziv Goldfeld, Kristjan Greenewald, Kengo Kato
Requests for name changes in the electronic proceedings will be accepted with 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.

**********
Online Robust Regression via SGD on the l1 loss
Scott Pesme, Nicolas Flammarion
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
PRANK: motion Prediction based on RANKing
Yuriy Biktairov, Maxim Stebelev, Irina Rudenko, Oleh Shliazhko, Boris Yangel
Predicting the motion of agents such as pedestrians or human-driven vehicles is one of the most critical problems in the autonomous driving domain. The overall safety of driving and the comfort of a passenger directly depend on its successful solution. The motion prediction problem also remains one of the most challenging problems in autonomous driving engineering, mainly due to high variance of the possible agent’s future behavior given a situation. The two phenomena responsible for the said variance are the multimodality caused by the uncertainty of the agent’s intent (e.g., turn right or move forward) and uncertainty in the realization of a given intent (e.g., which lane to turn into). To be useful within a real-time autonomous driving pipeline, a motion prediction system must provide efficient ways to describe and quantify this uncertainty, such as computing posterior modes and their probabilities or estimating density at the point corresponding to a given trajectory. It also should not put substantial density on physically impossible trajectories, as they can confuse the system processing the predictions. In this paper, we introduce the PRANK method, which satisfies these requirements. PRANK takes rasterized bird-eye images of agent’s surroundings as an input and extracts features of the scene with a convolutional neural network. It then produces the conditional distribution of agent’s trajectories plausible in the given scene. The key contribution of PRANK is a way to represent that distribution using nearest-neighbor methods in latent trajectory space, which allows for efficient inference in real time. We evaluate PRANK on the in-house and Argoverse datasets, where it shows competitive results.
**********
Fighting Copycat Agents in Behavioral Cloning from Observation Histories
Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, Yang Gao
Imitation learning trains policies to map from input observations to the actions that an expert would choose. In this setting, distribution shift frequently exacerbates the effect of misattributing expert actions to nuisance correlates among the observed variables. We observe that a common instance of this causal confusion occurs in partially observed settings when expert actions are strongly correlated over time: the imitator learns to cheat by predicting the expert's previous action, rather than the next action. To combat this "copycat problem", we propose an adversarial approach to learn a feature representation that removes excess information about the previous expert action nuisance correlate, while retaining the information necessary to predict the next action. In our experiments, our approach improves performance significantly across a variety of partially observed imitation learning tasks.
**********
Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model
Raphaël Berthier, Francis Bach, Pierre Gaillard
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Structured Prediction for Conditional Meta-Learning
Ruohan Wang, Yiannis Demiris, Carlo Ciliberto
The goal of optimization-based meta-learning is to find a single initialization shared across a distribution of tasks to speed up the process of learning new ta

sks. Conditional meta-learning seeks task-specific initialization to better capt
ure complex task distributions and improve performance. However, many existing c
onditional methods are difficult to generalize and lack theoretical guarantees.
In this work, we propose a new perspective on conditional meta-learning via stru
ctured prediction. We derive task-adaptive structured meta-learning (TASML), a p
rincipled framework that yields task-specific objective functions by weighing me
ta-training data on target tasks. Our non-parametric approach is model-agnostic
and can be combined with existing meta-learning methods to achieve conditioning.
 Empirically, we show that TASML improves the performance of existing meta-learn
ing models, and outperforms the state-of-the-art on benchmark datasets.
**********

Optimal Lottery Tickets via Subset Sum: Logarithmic Over-Parameterization is Suf
ficient
Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, Dimitris Papaili
opoulos
Requests for name changes in the electronic proceedings will be accepted with 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 Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes
Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Prat
ik Ringshia, Davide Testuggine
This work proposes a new challenge set for multimodal classification, focusing o
n
detecting hate speech in multimodal memes. It is constructed such that unimodal
models struggle and only multimodal models can succeed: difficult examples
(“benign confounders”) are added to the dataset to make it hard to rely on unimo
dal
signals. The task requires subtle reasoning, yet is straightforward to evaluate
as a binary classification problem. We provide baseline performance numbers
for unimodal models, as well as for multimodal models with various degrees of
sophistication. We find that state-of-the-art methods perform poorly compared to
humans, illustrating the difficulty of the task and highlighting the challenge t
hat this important problem poses to the community.
**********

Stochasticity of Deterministic Gradient Descent: Large Learning Rate for Multisc
ale Objective Function
Lingkai Kong, Molei Tao
This article suggests that deterministic Gradient Descent, which does not use an
y stochastic gradient approximation, can still exhibit stochastic behaviors. In
particular, it shows that if the objective function exhibit multiscale behaviors
, then in a large learning rate regime which only resolves the macroscopic but n
ot the microscopic details of the objective, the deterministic GD dynamics can b
ecome chaotic and convergent not to a local minimizer but to a statistical distr
ibution. In this sense, deterministic GD resembles stochastic GD even though no
stochasticity is injected. A sufficient condition is also established for approx
imating this long-time statistical limit by a rescaled Gibbs distribution, which
 for example allows escapes from local minima to be quantified. Both theoretical
 and numerical demonstrations are provided, and the theoretical part relies on t
he construction of a stochastic map that uses bounded noise (as opposed to Gauss
ian noise).
**********

Identifying Learning Rules From Neural Network Observables
Aran Nayebi, Sanjana Srivastava, Surya Ganguli, Daniel L. Yamins
The brain modifies its synaptic strengths during learning in order to better ada
pt to its environment. However, the underlying plasticity rules that govern lear
ning are unknown. Many proposals have been suggested, including Hebbian mechanis
ms, explicit error backpropagation, and a variety of alternatives. It is an open
 question as to what specific experimental measurements would need to be made to

determine whether any given learning rule is operative in a real biological system. In this work, we take a "virtual experimental" approach to this problem. Simulating idealized neuroscience experiments with artificial neural networks, we generate a large-scale dataset of learning trajectories of aggregate statistics measured in a variety of neural network architectures, loss functions, learning rule hyperparameters, and parameter initializations. We then take a discriminative approach, training linear and simple non-linear classifiers to identify learning rules from features based on these observables. We show that different classes of learning rules can be separated solely on the basis of aggregate statistics of the weights, activations, or instantaneous layer-wise activity changes, and that these results generalize to limited access to the trajectory and held-out architectures and learning curricula. We identify the statistics of each observable that are most relevant for rule identification, finding that statistics from network activities across training are more robust to unit undersampling and measurement noise than those obtained from the synaptic strengths. Our results suggest that activation patterns, available from electrophysiological recordings of post-synaptic activities on the order of several hundred units, frequently measured at wider intervals over the course of learning, may provide a good basis on which to identify learning rules.
**********

Optimal Approximation - Smoothness Tradeoffs for Soft-Max Functions
Alessandro Epasto, Mohammad Mahdian, Vahab Mirrokni, Emmanouil Zampetakis
respect to the Renyi Divergence, which provides improved theoretical and practical results in differentially private submodular optimization.
**********

Weakly-Supervised Reinforcement Learning for Controllable Behavior
Lisa Lee, Ben Eysenbach, Russ R. Salakhutdinov, Shixiang (Shane) Gu, Chelsea Finn
Reinforcement learning (RL) is a powerful framework for learning to take actions to solve tasks. However, in many settings, an agent must winnow down the inconceivably large space of all possible tasks to the single task that it is currently being asked to solve. Can we instead constrain the space of tasks to those that are semantically meaningful? In this work, we introduce a framework for using weak supervision to automatically disentangle this semantically meaningful subspace of tasks from the enormous space of nonsensical "chaff" tasks. We show that this learned subspace enables efficient exploration and provides a representation that captures distance between states. On a variety of challenging, vision-based continuous control problems, our approach leads to substantial performance gains, particularly as the complexity of the environment grows.
**********

Improving Policy-Constrained Kidney Exchange via Pre-Screening
Duncan McElfresh, Michael Curry, Tuomas Sandholm, John Dickerson
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  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 abstract structure for drawing by efficient motor program induction
Lucas Tian, Kevin Ellis, Marta Kryven, Josh Tenenbaum
Humans flexibly solve new problems that differ from those previously practiced. This ability to flexibly generalize is supported by learned concepts that represent useful structure common across different problems. Here we develop a naturalistic drawing task to study how humans rapidly acquire structured prior knowledge. The task requires drawing visual figures that share underlying structure, based on a set of composable geometric rules and simple objects. We show that people spontaneously learn abstract drawing procedures that support generalization, and propose a model of how learners can discover these reusable drawing procedures. Trained in the same setting as humans, and constrained to produce efficient motor actions, this model discovers new drawing program subroutines that generalize to test figures and resemble learned features of human behavior. These result

s suggest that two principles guiding motor program induction in the model - abs
traction (programs can reflect high-level structure that ignores figure-specific
 details) and compositionality (new programs are discovered by recombining previ
ously learned programs) - are key for explaining how humans learn structured int
ernal representations that guide flexible reasoning and learning.
**********

Why Do Deep Residual Networks Generalize Better than Deep Feedforward Networks?
--- A Neural Tangent Kernel Perspective

Kaixuan Huang, Yuqing Wang, Molei Tao, Tuo Zhao

Deep residual networks (ResNets) have demonstrated better generalization perform
ance than deep feedforward networks (FFNets). However, the theory behind such a
phenomenon is still largely unknown. This paper studies this fundamental problem
 in deep learning from a so-called ``neural tangent kernel'' perspective. Specif
ically, we first show that under proper conditions, as the width goes to infinit
y, training deep ResNets can be viewed as learning reproducing kernel functions
with some kernel function. We then compare the kernel of deep ResNets with that
of deep FFNets and discover that the class of functions induced by the kernel of
 FFNets is asymptotically not learnable, as the depth goes to infinity. In contr
ast, the class of functions induced by the kernel of ResNets does not exhibit su
ch degeneracy. Our discovery partially justifies the advantages of deep ResNets
over deep FFNets in generalization abilities. Numerical results are provided to
support our claim.
**********

Dual Instrumental Variable Regression

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, Anant Raj

We present a novel algorithm for non-linear instrumental variable (IV) regressio
n, DualIV, which simplifies traditional two-stage methods via a dual formulation
. Inspired by problems in stochastic programming, we show that two-stage procedu
res for non-linear IV regression can be reformulated as a convex-concave saddle-
point problem. Our formulation enables us to circumvent the first-stage regressi
on which is a potential bottleneck in real-world applications. We develop a simp
le kernel-based algorithm with an analytic solution based on this formulation. E
mpirical results show that we are competitive to existing, more complicated algo
rithms for non-linear instrumental variable regression.
**********

Stochastic Gradient Descent in Correlated Settings: A Study on Gaussian Processe
s

Hao Chen, Lili Zheng, Raed AL Kontar, Garvesh Raskutti

Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Interventional Few-Shot Learning

Zhongqi Yue, Hanwang Zhang, Qianru Sun, Xian-Sheng Hua

We uncover an ever-overlooked deficiency in the prevailing Few-Shot Learning (FS
L) methods: the pre-trained knowledge is indeed a confounder that limits the per
formance. This finding is rooted from our causal assumption: a Structural Causal
 Model (SCM) for the causalities among the pre-trained knowledge, sample feature
s, and labels. Thanks to it, we propose a novel FSL paradigm: Interventional Few
-Shot Learning (IFSL). Specifically, we develop three effective IFSL algorithmic
 implementations based on the backdoor adjustment, which is essentially a causal
 intervention towards the SCM of many-shot learning: the upper-bound of FSL in a
 causal view. It is worth noting that the contribution of IFSL is orthogonal to
existing fine-tuning and meta-learning based FSL methods, hence IFSL can improve
 all of them, achieving a new 1-/5-shot state-of-the-art on miniImageNet, tiered
ImageNet, and cross-domain CUB. Code is released at https://github.com/yue-zhong
qi/ifsl.
**********

Minimax Value Interval for Off-Policy Evaluation and Policy Optimization

Nan Jiang, Jiawei Huang
We study minimax methods for off-policy evaluation (OPE) using value functions a nd marginalized importance weights. Despite that they hold promises of overcomin g the exponential variance in traditional importance sampling, several key probl ems remain:
(1) They require function approximation and are generally biased. For the sake o f trustworthy OPE, is there anyway to quantify the biases?
(2) They are split into two styles (“weight-learning” vs “value-learning”). Can we unify them?
In this paper we answer both questions positively. By slightly altering the deri vation of previous methods (one from each style), we unify them into a single va lue interval that comes with a special type of double robustness: when either th e value-function or the importance-weight class is well specified, the interval is valid and its length quantifies the misspecification of the other class. Our interval also provides a unified view of and new insights to some recent methods , and we further explore the implications of our results on exploration and expl oitation in off-policy policy optimization with insufficient data coverage.
**********

Biased Stochastic First-Order Methods for Conditional Stochastic Optimization an d Applications in Meta Learning
Yifan Hu, Siqi Zhang, Xin Chen, Niao He
Conditional stochastic optimization covers a variety of applications ranging fro m invariant learning and causal inference to meta-learning. However, constructin g unbiased gradient estimators for such problems is challenging due to the compo sition structure. As an alternative, we propose a biased stochastic gradient des cent (BSGD) algorithm and study the bias-variance tradeoff under different struc tural assumptions. We establish the sample complexities of BSGD for strongly con vex, convex, and weakly convex objectives under smooth and non-smooth conditions . Our lower bound analysis shows that the sample complexities of BSGD cannot be improved for general convex objectives and nonconvex objectives except for smoot h nonconvex objectives with Lipschitz continuous gradient estimator.  For this s pecial setting, we propose an accelerated algorithm called biased SpiderBoost (B SpiderBoost) that matches the lower bound complexity.  We further conduct numeri cal experiments on invariant logistic regression and model-agnostic meta-learnin g to illustrate the performance of BSGD and BSpiderBoost.
**********

ShiftAddNet: A Hardware-Inspired Deep Network
Haoran You, Xiaohan Chen, Yongan Zhang, Chaojian Li, Sicheng Li, Zihao Liu, Zhan gyang Wang, Yingyan Lin
Multiplication (e.g., convolution) is arguably a cornerstone of modern deep neur al networks (DNNs). However, intensive multiplications cause expensive resource costs that challenge DNNs' deployment on resource-constrained edge devices, driv ing several attempts for multiplication-less deep networks. This paper presented  ShiftAddNet, whose main inspiration is drawn from a common practice in energy-e fficient hardware implementation, that is,  multiplication can be instead perfor med with additions and logical bit-shifts. We leverage this idea to explicitly p arameterize deep networks in this way, yielding a new type of deep network that involves only bit-shift and additive weight layers. This hardware-inspired Shift AddNet immediately leads to both energy-efficient inference and training, withou t compromising the expressive capacity compared to standard DNNs. The two comple mentary operation types (bit-shift and add) additionally enable finer-grained co ntrol of the model's learning capacity, leading to more flexible trade-off betwe en accuracy and (training) efficiency, as well as improved robustness to quantiz ation and pruning. We conduct extensive experiments and ablation studies, all ba cked up by our FPGA-based ShiftAddNet implementation and energy measurements. Co mpared to existing DNNs or other multiplication-less models, ShiftAddNet aggress ively reduces over 80% hardware-quantified energy cost of DNNs training and infe rence, while offering comparable or better accuracies. Codes and pre-trained mod els are available at https://github.com/RICE-EIC/ShiftAddNet.
**********

Network-to-Network Translation with Conditional Invertible Neural Networks
Robin Rombach, Patrick Esser, Bjorn Ommer
Given the ever-increasing computational costs of modern machine learning models, we need to find new ways to reuse such expert models and thus tap into the resources that have been invested in their creation. Recent work suggests that the power of these massive models is captured by the representations they learn. Therefore, we seek a model that can relate between different existing representations and propose to solve this task with a conditionally invertible network. This network demonstrates its capability by (i) providing generic transfer between diverse domains, (ii) enabling controlled content synthesis by allowing modification in other domains, and (iii) facilitating diagnosis of existing representations by translating them into interpretable domains such as images. Our domain transfer network can translate between fixed representations without having to learn or finetune them. This allows users to utilize various existing domain-specific expert models from the literature that had been trained with extensive computational resources. Experiments on diverse conditional image synthesis tasks, competitive image modification results and experiments on image-to-image and text-to-image generation demonstrate the generic applicability of our approach. For example, we translate between BERT and BigGAN, state-of-the-art text and image models to provide text-to-image generation, which neither of both experts can perform on their own.
**********
Intra-Processing Methods for Debiasing Neural Networks
Yash Savani, Colin White, Naveen Sundar Govindarajulu
As deep learning models become tasked with more and more decisions that impact human lives, such as criminal recidivism, loan repayment, and face recognition for law enforcement, bias is becoming a growing concern. Debiasing algorithms are typically split into three paradigms: pre-processing, in-processing, and post-processing. However, in computer vision or natural language applications, it is common to start with a large generic model and then fine-tune to a specific use-case. Pre- or in-processing methods would require retraining the entire model from scratch, while post-processing methods only have black-box access to the model, so they do not leverage the weights of the trained model. Creating debiasing algorithms specifically for this fine-tuning use-case has largely been neglected.
**********
Finding Second-Order Stationary Points Efficiently in Smooth Nonconvex Linearly Constrained Optimization Problems
Songtao Lu, Meisam Razaviyayn, Bo Yang, Kejun Huang, Mingyi Hong
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  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-based Policy Optimization with Unsupervised Model Adaptation
Jian Shen, Han Zhao, Weinan Zhang, Yong Yu
Model-based reinforcement learning methods learn a dynamics model with real data sampled from the environment and leverage it to generate simulated data to derive an agent. However, due to the potential distribution mismatch between simulated data and real data, this could lead to degraded performance. Despite much effort being devoted to reducing this distribution mismatch, existing methods fail to solve it explicitly. In this paper, we investigate how to bridge the gap between real and simulated data due to inaccurate model estimation for better policy optimization. To begin with, we first derive a lower bound of the expected return, which naturally inspires a bound maximization algorithm by aligning the simulated and real data distributions. To this end, we propose a novel model-based reinforcement learning framework AMPO, which introduces unsupervised model adaptation to minimize the integral probability metric (IPM) between feature distributions from real and simulated data. Instantiating our framework with Wasserstein-1 distance gives a practical model-based approach. Empirically, our approach achieves state-of-the-art performance in terms of sample efficiency on a range of c

ontinuous control benchmark tasks.
**********
Implicit Regularization and Convergence for Weight Normalization
Xiaoxia Wu, Edgar Dobriban, Tongzheng Ren, Shanshan Wu, Zhiyuan Li, Suriya Gunasekar, Rachel Ward, 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.
**********
Geometric All-way Boolean Tensor Decomposition
Changlin Wan, Wennan Chang, Tong Zhao, Sha Cao, Chi Zhang
Boolean tensor has been broadly utilized in representing high dimensional logical data collected on spatial, temporal and/or other relational domains. Boolean Tensor Decomposition (BTD) factorizes a binary tensor into the Boolean sum of multiple rank-1 tensors, which is an NP-hard problem. Existing BTD methods have been limited by their high computational cost, in applications to large scale or higher order tensors. In this work, we presented a computationally efficient BTD algorithm, namely Geometric Expansion for all-order Tensor Factorization (GETF), that sequentially identifies the rank-1 basis components for a tensor from a geometric perspective. We conducted rigorous theoretical analysis on the validity as well as algorithemic efficiency of GETF in decomposing all-order tensor. Experiments on both synthetic and real-world data demonstrated that GETF has significantly improved performance in reconstruction accuracy, extraction of latent structures and it is an order of magnitude faster than other state-of-the-art methods.
**********
Modular Meta-Learning with Shrinkage
Yutian Chen, Abram L. Friesen, Feryal Behbahani, Arnaud Doucet, David Budden, Matthew Hoffman, Nando de Freitas
Many real-world problems, including multi-speaker text-to-speech synthesis, can greatly benefit from the ability to meta-learn large models with only a few task-specific components. Updating only these task-specific modules then allows the model to be adapted to low-data tasks for as many steps as necessary without risking overfitting. Unfortunately, existing meta-learning methods either do not scale to long adaptation or else rely on handcrafted task-specific architectures. Here, we propose a meta-learning approach that obviates the need for this often sub-optimal hand-selection. In particular, we develop general techniques based on Bayesian shrinkage to automatically discover and learn both task-specific and general reusable modules. Empirically, we demonstrate that our method discovers a small set of meaningful task-specific modules and outperforms existing meta-learning approaches in domains like few-shot text-to-speech that have little task data and long adaptation horizons. We also show that existing meta-learning methods including MAML, iMAML, and Reptile emerge as special cases of our method.
**********
A/B Testing in Dense Large-Scale Networks: Design and Inference
Preetam Nandy, Kinjal Basu, Shaunak Chatterjee, Ye Tu
Design of experiments and estimation of treatment effects in large-scale networks, in the presence of strong interference, is a challenging and important problem. Most existing methods' performance deteriorates as the density of the network increases. In this paper, we present a novel strategy for accurately estimating the causal effects of a class of treatments in a dense large-scale network. First, we design an approximate randomized controlled experiment by solving an optimization problem to allocate treatments in the presence of competition among neighboring nodes. Then we apply an importance sampling adjustment to correct for any leftover bias (from the approximation) in estimating average treatment effects. We provide theoretical guarantees, verify robustness in a simulation study, and validate the scalability and usefulness of our procedure in a real-world experiment on a large social network.
**********

What Neural Networks Memorize and Why: Discovering the Long Tail via Influence E
stimation
Vitaly Feldman, Chiyuan Zhang
Deep learning algorithms are well-known to have a propensity for fitting the tra
ining data very well and often fit even outliers and mislabeled data points. Suc
h fitting requires memorization of training data labels, a phenomenon that has a
ttracted significant research interest but has not been given a compelling expla
nation so far. A recent work of Feldman (2019) proposes a theoretical explanatio
n for this phenomenon based on a combination of two insights. First, natural ima
ge and data distributions are (informally) known to be long-tailed, that is have
 a significant fraction of rare and atypical examples. Second, in a simple theor
etical model such memorization is necessary for achieving close-to-optimal gener
alization error when the data distribution is long-tailed. However, no direct em
pirical evidence for this explanation or even an approach for obtaining such evi
dence were given.
**********

Partially View-aligned Clustering
Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, Xi Peng
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Partial Optimal Tranport with applications on Positive-Unlabeled Learning
Laetitia Chapel, Mokhtar Z. Alaya, Gilles Gasso
Classical optimal transport  problem  seeks a transportation map that preserves
the total mass between two probability distributions, requiring their masses to
be equal. This  may be too restrictive in some applications such as color or sha
pe matching, since the distributions may have arbitrary masses and/or only a fra
ction of the total mass has to be transported. In this paper, we address the par
tial Wasserstein and Gromov-Wasserstein problems and propose exact algorithms to
 solve them. We showcase the new formulation in a positive-unlabeled (PU) learni
ng application. To the best of our knowledge, this is the first application of o
ptimal transport in this context and we first highlight that partial Wasserstein
-based metrics prove effective in usual PU learning settings. We then demonstrat
e that partial Gromov-Wasserstein metrics are efficient in scenarii in which the
 samples from the positive and the unlabeled datasets come from different domain
s or have different features.
**********

Toward the Fundamental Limits of Imitation Learning
Nived Rajaraman, Lin Yang, Jiantao Jiao, Kannan Ramchandran
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Logarithmic Pruning is All You Need
Laurent Orseau, Marcus Hutter, Omar Rivasplata
The Lottery Ticket Hypothesis is a conjecture that every large neural network co
ntains a subnetwork that, when trained in isolation, achieves comparable perform
ance to the large network.
An even stronger conjecture has been proven recently: Every sufficiently overpar
ameterized network contains a subnetwork that, even without training, achieves c
omparable accuracy to the trained large network.
This theorem, however, relies on a number of strong assumptions and guarantees a
 polynomial factor on the size of the large network compared to the target funct
ion.
In this work, we remove the most limiting assumptions of this previous work whil
e providing significantly tighter bounds:
the overparameterized network only needs a logarithmic factor (in all variables

but depth) number of neurons per weight of the target subnetwork.
**********
Hold me tight! Influence of discriminative features on deep network boundaries
Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi, Pascal Frossard
Important insights towards the explainability of neural networks reside in the characteristics of their decision boundaries. In this work, we borrow tools from the field of adversarial robustness, and propose a new perspective that relates dataset features to the distance of samples to the decision boundary. This enables us to carefully tweak the position of the training samples and measure the induced changes on the boundaries of CNNs trained on large-scale vision datasets. We use this framework to reveal some intriguing properties of CNNs. Specifically, we rigorously confirm that neural networks exhibit a high invariance to non-discriminative features, and show that the decision boundaries of a DNN can only exist as long as the classifier is trained with some features that hold them together. Finally, we show that the construction of the decision boundary is extremely sensitive to small perturbations of the training samples, and that changes in certain directions can lead to sudden invariances in the orthogonal ones. This is precisely the mechanism that adversarial training uses to achieve robustness.
**********
Learning from Mixtures of Private and Public Populations
Raef Bassily, Shay Moran, Anupama Nandi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Weight Perturbation Helps Robust Generalization
Dongxian Wu, Shu-Tao Xia, Yisen Wang
The study on improving the robustness of deep neural networks against adversarial examples grows rapidly in recent years. Among them, adversarial training is the most promising one, which flattens the \textit{input loss landscape} (loss change with respect to input) via training on adversarially perturbed examples. However, how the widely used \textit{weight loss landscape} (loss change with respect to weight) performs in adversarial training is rarely explored. In this paper, we investigate the weight loss landscape from a new perspective, and identify a clear correlation between the flatness of weight loss landscape and robust generalization gap. Several well-recognized adversarial training improvements, such as early stopping, designing new objective functions, or leveraging unlabeled data, all implicitly flatten the weight loss landscape. Based on these observations, we propose a simple yet effective \textit{Adversarial Weight Perturbation (AWP)} to explicitly regularize the flatness of weight loss landscape, forming a \textit{double-perturbation} mechanism in the adversarial training framework that adversarially perturbs both inputs and weights. Extensive experiments demonstrate that AWP indeed brings flatter weight loss landscape and can be easily incorporated into various existing adversarial training methods to further boost their adversarial robustness.
**********
Stateful Posted Pricing with Vanishing Regret via Dynamic Deterministic Markov Decision Processes
Yuval Emek, Ron Lavi, Rad Niazadeh, Yangguang Shi
In this paper, a rather general online problem called \emph{dynamic resource allocation with capacity constraints (DRACC)} is introduced and studied in the realm of posted price mechanisms. This problem subsumes several applications of stateful pricing, including but not limited to posted prices for online job scheduling and matching over a dynamic bipartite graph. As the existing online learning techniques do not yield vanishing-regret mechanisms for this problem, we develop a novel online learning framework defined over deterministic Markov decision processes with \emph{dynamic} state transition and reward functions. We then prove that if the Markov decision process is guaranteed to admit an oracle that can simulate any given policy from any initial state with bounded loss --- a conditio

n that is satisfied in the DRACC problem --- then the online learning problem ca
n be solved with vanishing regret. Our proof technique is based on a reduction t
o online learning with \emph{switching cost}, in which an online decision maker
incurs an extra cost every time she switches from one arm to another. We formall
y demonstrate this connection and further show how DRACC can be used in our prop
osed applications of stateful pricing.
**********

Adversarial Self-Supervised Contrastive Learning
Minseon Kim, Jihoon Tack, Sung Ju Hwang
Existing adversarial learning approaches mostly use class labels to generate adv
ersarial samples that lead to incorrect predictions, which are then used to augm
ent the training of the model for improved robustness. While some recent works p
ropose semi-supervised adversarial learning methods that utilize unlabeled data,
 they still require class labels. However, do we really need class labels at all
, for adversarially robust training of deep neural networks? In this paper, we p
ropose a novel adversarial attack for unlabeled data, which makes the model conf
use the instance-level identities of the perturbed data samples. Further, we pre
sent a self-supervised contrastive learning framework to adversarially train a r
obust neural network without labeled data, which aims to maximize the similarity
 between a random augmentation of a data sample and its instance-wise adversaria
l perturbation. We validate our method, Robust Contrastive Learning (RoCL), on m
ultiple benchmark datasets, on which it obtains comparable robust accuracy over
state-of-the-art supervised adversarial learning methods, and significantly impr
oved robustness against the \emph{black box} and unseen types of attacks. Moreov
er, with further joint fine-tuning with supervised adversarial loss, RoCL obtain
s even higher robust accuracy over using self-supervised learning alone. Notably
, RoCL also demonstrate impressive results in robust transfer learning.
**********

Normalizing Kalman Filters for Multivariate Time Series Analysis
Emmanuel de Bézenac, Syama Sundar Rangapuram, Konstantinos Benidis, Michael Bohl
ke-Schneider, Richard Kurle, Lorenzo Stella, Hilaf Hasson, Patrick Gallinari, Ti
m Januschowski
This paper tackles the modelling of large, complex and multivariate time series
panels in a probabilistic setting. To this extent, we present a novel approach r
econciling classical state space models with deep learning methods. By augmentin
g state space models with normalizing flows, we mitigate imprecisions stemming f
rom idealized assumptions in state space models. The resulting model is highly f
lexible while still retaining many of the attractive properties of state space m
odels, e.g., uncertainty and observation errors are properly accounted for, infe
rence is tractable, sampling is efficient, good generalization performance is ob
served, even in low data regimes. We demonstrate competitiveness against state-o
f-the-art deep learning methods on the tasks of forecasting real world data and
handling varying levels of missing data.
**********

Learning to summarize with human feedback
Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss
, Alec Radford, Dario Amodei, Paul F. Christiano
As language models become more powerful, training and evaluation are increasingl
y bottlenecked by the data and metrics used for a particular task.  For example,
 summarization models are often trained to predict human reference summaries and
 evaluated using ROUGE, but both of these metrics are rough proxies for what we
really care about---summary quality. In this work, we show that it is possible t
o significantly improve summary quality by training a model to optimize for huma
n preferences.  We collect a large, high-quality dataset of human comparisons be
tween summaries, train a model to predict the human-preferred summary, and use t
hat model as a reward function to fine-tune a summarization policy using reinfor
cement learning. We apply our method to a version of the TL;DR dataset of Reddit
 posts and find that our models significantly outperform both human reference su
mmaries and much larger models fine-tuned with supervised learning alone. Our mo
dels also transfer to CNN/DM news articles, producing summaries nearly as good a

s the human reference without any news-specific fine-tuning.  We conduct extensi
ve analyses to understand our human feedback dataset and fine-tuned models.  We
establish that our reward model generalizes to new datasets, and that optimizing
 our reward model results in better summaries than optimizing ROUGE according to
 humans.  We hope the evidence from our paper motivates machine learning researc
hers to pay closer attention to how their training loss affects the model behavi
or they actually want.
**********

Fourier Spectrum Discrepancies in Deep Network Generated Images
Tarik Dzanic, Karan Shah, Freddie Witherden
Advancements in deep generative models such as generative adversarial networks a
nd variational autoencoders have resulted in the ability to generate realistic i
mages that are visually indistinguishable from real images which raises concerns
 about their potential malicious usage. In this paper, we present an analysis of
 the high-frequency Fourier modes of real and deep network generated images and
show that deep network generated images share an observable, systematic shortcom
ing in replicating the attributes of these high-frequency modes. Using this, we
propose a novel detection method based on the frequency spectrum of the images w
hich is able to achieve an accuracy of up to 99.2% in classifying real and deep
network generated images from various GAN and VAE architectures on a dataset of
5000 images with as few as 8 training examples. Furthermore, we show the impact
of image transformations such as compression, cropping, and resolution reduction
 on the classification accuracy and suggest a method for modifying the high-freq
uency attributes of deep network generated images to mimic real images.
**********

Lamina-specific neuronal properties promote robust, stable signal propagation in
 feedforward networks
Dongqi Han, Erik De Schutter, Sungho Hong
Feedforward networks (FFN) are ubiquitous structures in neural systems and have
been studied to understand mechanisms of reliable signal and information transmi
ssion. In many FFNs, neurons in one layer have intrinsic properties that are dis
tinct from those in their pre-/postsynaptic layers, but how this affects network
-level information processing remains unexplored. Here we show that layer-to-lay
er heterogeneity arising from lamina-specific cellular properties facilitates si
gnal and information transmission in FFNs. Specifically, we found that signal tr
ansformations, made by each layer of neurons on an input-driven spike signal, de
modulate signal distortions introduced by preceding layers. This mechanism boost
s information transfer carried by a propagating spike signal, and thereby suppor
ts reliable spike signal and information transmission in a deep FFN. Our study s
uggests that distinct cell types in neural circuits, performing different comput
ational functions, facilitate information processing on the whole.
**********

Learning Dynamic Belief Graphs to Generalize on Text-Based Games
Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoi
ne Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, Will Hami
lton
Playing text-based games requires skills in processing natural language and sequ
ential decision making. Achieving human-level performance on text-based games re
mains an open challenge, and prior research has largely relied on hand-crafted s
tructured representations and heuristics. In this work, we investigate how an ag
ent can plan and generalize in text-based games using graph-structured represent
ations learned end-to-end from raw text. We propose a novel graph-aided transfor
mer agent (GATA) that infers and updates latent belief graphs during planning to
 enable effective action selection by capturing the underlying game dynamics. GA
TA is trained using a combination of reinforcement and self-supervised learning.
 Our work demonstrates that the learned graph-based representations help agents
converge to better policies than their text-only counterparts and facilitate eff
ective generalization across game configurations. Experiments on 500+ unique gam
es from the TextWorld suite show that our best agent outperforms text-based base
lines by an average of 24.2%.

**********
Triple descent and the two kinds of overfitting: where & why do they appear?
Stéphane d'Ascoli, Levent Sagun, Giulio Biroli
A recent line of research has highlighted the existence of a ``double descent'' phenomenon in deep learning, whereby increasing the number of training examples N causes the generalization error of neural networks to peak when N is of the same order as the number of parameters P. In earlier works, a similar phenomenon was shown to exist in simpler models such as linear regression, where the peak instead occurs when N is equal to the input dimension D. Since both peaks coincide with the interpolation threshold, they are often conflated in the litterature. In this paper, we show that despite their apparent similarity, these two scenarios are inherently different. In fact, both peaks can co-exist when neural networks are applied to noisy regression tasks. The relative size of the peaks is then governed by the degree of nonlinearity of the activation function. Building on recent developments in the analysis of random feature models, we provide a theoretical ground for this sample-wise triple descent. As shown previously, the nonlinear peak at N=P is a true divergence caused by the extreme sensitivity of the output function to both the noise corrupting the labels and the initialization of the random features (or the weights in neural networks). This peak survives in the absence of noise, but can be suppressed by regularization. In contrast, the linear peak at N=D is solely due to overfitting the noise in the labels, and forms earlier during training. We show that this peak is implicitly regularized by the nonlinearity, which is why it only becomes salient at high noise and is weakly affected by explicit regularization.
Throughout the paper, we compare the analytical results obtained in the random feature model with the outcomes of numerical experiments involving realistic neural networks.
**********
Multimodal Graph Networks for Compositional Generalization in Visual Question Answering
Raeid Saqur, Karthik Narasimhan
Compositional generalization is a key challenge in grounding natural language to visual perception. While deep learning models have achieved great success in multimodal tasks like visual question answering, recent studies have shown that they fail to generalize to new inputs that are simply an unseen combination of those seen in the training distribution. In this paper, we propose to tackle this challenge by employing neural factor graphs to induce a tighter coupling between concepts in different modalities (e.g. images and text). Graph representations are inherently compositional in nature and allow us to capture entities, attributes and relations in a scalable manner. Our model first creates a multimodal graph, processes it with a graph neural network to induce a factor correspondence matrix, and then outputs a symbolic program to predict answers to questions. Empirically, our model achieves close to perfect scores on a caption truth prediction problem and  state-of-the-art results on the recently introduced CLOSURE dataset, improving on the mean overall accuracy across seven compositional templates by 4.77\% over previous approaches.
**********
Learning Graph Structure With A Finite-State Automaton Layer
Daniel Johnson, Hugo Larochelle, Daniel Tarlow
Graph-based neural network models are producing strong results in a number of domains, in part because graphs provide flexibility to encode domain knowledge in the form of relational structure (edges) between nodes in the graph. In practice, edges are used both to represent intrinsic structure (e.g., abstract syntax trees of programs) and more abstract relations that aid reasoning for a downstream task (e.g., results of relevant program analyses). In this work, we study the problem of learning to derive abstract relations from the intrinsic graph structure. Motivated by their power in program analyses, we consider relations defined by paths on the base graph accepted by a finite-state automaton. We show how to learn these relations end-to-end by relaxing the problem into learning finite-state automata policies on a graph-based POMDP and then training these policies us

ing implicit differentiation. The result is a differentiable Graph Finite-State Automaton (GFSA) layer that adds a new edge type (expressed as a weighted adjacency matrix) to a base graph. We demonstrate that this layer can find shortcuts in grid-world graphs and reproduce simple static analyses on Python programs. Additionally, we combine the GFSA layer with a larger graph-based model trained end-to-end on the variable misuse program understanding task, and find that using the GFSA layer leads to better performance than using hand-engineered semantic edges or other baseline methods for adding learned edge types.
**********
A Universal Approximation Theorem of Deep Neural Networks for Expressing Probability Distributions
Yulong Lu, Jianfeng Lu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Unsupervised object-centric video generation and decomposition in 3D
Paul Henderson, Christoph H. Lampert
A natural approach to generative modeling of videos is to represent them as a composition of moving objects. Recent works model a set of 2D sprites over a slowly-varying background, but without considering the underlying 3D scene that gives rise to them. We instead propose to model a video as the view seen while moving through a scene with multiple 3D objects and a 3D background. Our model is trained from monocular videos without any supervision, yet learns to generate coherent 3D scenes containing several moving objects. We conduct detailed experiments on two datasets, going beyond the visual complexity supported by state-of-the-art generative approaches. We evaluate our method on depth-prediction and 3D object detection---tasks which cannot be addressed by those earlier works---and show it out-performs them even on 2D instance segmentation and tracking.
**********
Domain Generalization for Medical Imaging Classification with Linear-Dependency Regularization
Haoliang Li, Yufei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, Alex Kot
Recently, we have witnessed great progress in the field of medical imaging classification by adopting deep neural networks. However, the recent advanced models still require accessing sufficiently large and representative datasets for training, which is often unfeasible in clinically realistic environments. When trained on limited datasets, the deep neural network is lack of generalization capability, as the trained deep neural network on data within a certain distribution (e.g. the data captured by a certain device vendor or patient population) may not be able to generalize to the data with another distribution. In this paper, we introduce a simple but effective approach to improve the generalization capability of deep neural networks in the field of medical imaging classification. Motivated by the observation that the domain variability of the medical images is to some extent compact, we propose to learn a representative feature space through variational encoding with a novel linear-dependency regularization term to capture the shareable information among medical data collected from different domains. As a result, the trained neural network is expected to equip with better generalization capability to the ``unseen" medical data. Experimental results on two challenging medical imaging classification tasks indicate that our method can achieve better cross-domain generalization capability compared with state-of-the-art baselines.
**********
Multi-label classification: do Hamming loss and subset accuracy really conflict with each other?
Guoqiang Wu, Jun 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.
**********

A Novel Automated Curriculum Strategy to Solve Hard Sokoban Planning Instances
Dieqiao Feng, Carla P. Gomes, Bart Selman
In recent years, we have witnessed tremendous progress in deep reinforcement lea rning (RL) for tasks such as Go, Chess, video games, and robot control. Neverthe less, other combinatorial domains, such as AI planning, still pose considerable challenges for RL approaches. The key difficulty in those domains is that a posi tive reward signal becomes {\em exponentially rare} as the minimal solution leng th increases. So, an RL approach loses its training signal. There has been prom ising recent progress by using a curriculum-driven learning approach that is de signed to solve a single hard instance. We present a novel {\em automated} curr iculum approach that dynamically selects from a pool of unlabeled training insta nces of varying task complexity guided by our {\em difficulty quantum momentu m} strategy. We show how the smoothness of the task hardness impacts the fina l learning results. In particular, as the size of the instance pool increases, the ``hardness gap'' decreases, which facilitates a smoother automated curriculu m based learning process. Our automated curriculum approach dramatically improve s upon the previous approaches. We show our results on Sokoban, which is a tra ditional PSPACE-complete planning problem and presents a great challenge even fo r specialized solvers. Our RL agent can solve hard instances that are far out of reach for any previous state-of-the-art Sokoban solver. In particular, our appr oach can uncover plans that require hundreds of steps, while the best previous s earch methods would take many years of computing time to solve such instances. I n addition, we show that we can further boost the RL performance with an intric ate coupling of our automated curriculum approach with a curiosity-driven searc h strategy and a graph neural net representation.
**********

Causal analysis of Covid-19 Spread in Germany
Atalanti Mastakouri, Bernhard Schölkopf
In this work, we study the causal relations among German regions in terms of the spread of Covid-19 since the beginning of the pandemic, taking into account the restriction policies that were applied by the different federal states. We loos e a strictly formulated assumption for a causal feature selection method for tim e series data, robust to latent confounders, which we subsequently apply on Covi d-19 case numbers. We present findings about the spread of the virus in Germany and the causal impact of restriction measures, discussing the role of various po licies in containing the spread. Since our results are based on rather limited t arget time series (only the numbers of reported cases), care should be exercised in interpreting them. However, it is encouraging that already such limited data seems to contain causal signals. This suggests that as more data becomes availa ble, our causal approach may contribute towards meaningful causal analysis of po litical interventions on the development of Covid-19, and thus also towards the development of rational and data-driven methodologies for choosing interventions .
**********

Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms
Thomas Berrett, Cristina Butucea
We find separation rates for testing multinomial or more general discrete distri butions under the constraint of alpha-local differential privacy. We construct e fficient randomized algorithms and test procedures, in both the case where only non-interactive privacy mechanisms are allowed and also in the case where all se quentially interactive privacy mechanisms are allowed. The separation rates are faster in the latter case. We prove general information theoretical bounds that allow us to establish the optimality of our algorithms among all pairs of privac y mechanisms and test procedures, in most usual cases. Considered examples inclu de testing uniform, polynomially and exponentially decreasing distributions.
**********

Adaptive Gradient Quantization for Data-Parallel SGD

Fartash Faghri, Iman Tabrizian, Ilia Markov, Dan Alistarh, Daniel M. Roy, Ali Ramezani-Kebrya
Many communication-efficient variants of SGD use gradient quantization schemes. These schemes are often heuristic and fixed over the course of training. We empirically observe that the statistics of gradients of deep models change during the training. Motivated by this observation, we introduce two adaptive quantization schemes, ALQ and AMQ. In both schemes, processors update their compression schemes in parallel by efficiently computing sufficient statistics of a parametric distribution. We improve the validation accuracy by almost 2% on CIFAR-10 and 1% on ImageNet in challenging low-cost communication setups. Our adaptive methods are also significantly more robust to the choice of hyperparameters.
**********
Finite Continuum-Armed Bandits
Solenne Gaucher
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies
Itai Gat, Idan Schwartz, Alexander Schwing, Tamir Hazan
Many recent datasets contain a variety of different data modalities, for instance, image, question, and answer data in visual question answering (VQA). When training deep net classifiers on those multi-modal datasets, the modalities get exploited at different scales, i.e., some modalities can more easily contribute to the classification results than others. This is suboptimal because the classifier is inherently biased towards a subset of the modalities. To alleviate this shortcoming, we propose a novel regularization term based on the functional entropy. Intuitively, this term encourages to balance the contribution of each modality to the classification result. However, regularization with the functional entropy is challenging. To address this, we develop a method based on the log-Sobolev inequality, which bounds the functional entropy with the functional-Fisher-information. Intuitively, this maximizes the amount of information that the modalities contribute. On the two challenging multi-modal datasets VQA-CPv2, and SocialIQ, we obtain state-of-the-art results while more uniformly exploiting the modalities. In addition, we demonstrate the efficacy of our method on Colored MNIST.
**********
Compact task representations as a normative model for higher-order brain activity
Severin Berger, Christian K. Machens
Higher-order brain areas such as the frontal cortices are considered essential for the flexible solution of tasks. However, the precise computational role of these areas is still debated. Indeed, even for the simplest of tasks, we cannot really explain how the measured brain activity, which evolves over time in complicated ways, relates to the task structure. Here, we follow a normative approach, based on integrating the principle of efficient coding with the framework of Markov decision processes (MDP). More specifically, we focus on MDPs whose state is based on action-observation histories, and we show how to compress the state space such that unnecessary redundancy is eliminated, while task-relevant information is preserved. We show that the efficiency of a state space representation depends on the (long-term) behavioural goal of the agent, and we distinguish between model-based and habitual agents. We apply our approach to simple tasks that require short-term memory, and we show that the efficient state space representations reproduce the key dynamical features of recorded neural activity in frontal areas (such as ramping, sequentiality, persistence). If we additionally assume that neural systems are subject to accuracy-cost tradeoffs, we find a surprising match to neural data on a population level.
**********
Robust-Adaptive Control of Linear Systems: beyond Quadratic Costs

Edouard Leurent, Odalric-Ambrym Maillard, Denis Efimov
We consider the problem of robust and adaptive model predictive control (MPC) of a linear system, with unknown parameters that are learned along the way (adaptive), in a critical setting where failures must be prevented (robust). This problem has been studied from different perspectives by different communities. However, the existing theory deals only with the case of quadratic costs (the LQ problem), which limits applications to stabilisation and tracking tasks only. In order to handle more general (non-convex) costs that naturally arise in many practical problems, we carefully select and bring together several tools from different communities, namely non-asymptotic linear regression, recent results in interval prediction, and tree-based planning. Combining and adapting the theoretical guarantees at each layer is non trivial, and we provide the first end-to-end suboptimality analysis for this setting. Interestingly, our analysis naturally adapts to handle many models and combines with a data-driven robust model selection strategy, which enables to relax the modelling assumptions. Last, we strive to preserve tractability at any stage of the method, that we illustrate on two challenging simulated environments.
**********

Co-exposure Maximization in Online Social Networks
Sijing Tu, Cigdem Aslay, Aristides Gionis
Social media has created new ways for citizens to stay informed on societal matters and participate in political discourse. However, with its algorithmically-curated and virally-propagating content, social media has contributed further to the polarization of opinions by reinforcing users' existing viewpoints. An emerging line of research seeks to understand how content-recommendation algorithms can be re-designed to mitigate societal polarization amplified by social-media interactions. In this paper, we study the problem of allocating seed users to opposing campaigns: by drawing on the equal-time rule of political campaigning on traditional media, our goal is to allocate seed users to campaigners with the aim to maximize the expected number of users who are co-exposed to both campaigns.
**********

UCLID-Net: Single View Reconstruction in Object Space
Benoit Guillard, Edoardo Remelli, Pascal Fua
Most state-of-the-art deep geometric learning single-view reconstruction approaches rely on encoder-decoder architectures that output either shape parametrizations or implicit representations. However, these representations rarely preserve the Euclidean structure of the 3D space objects exist in. In this paper, we show that building a geometry preserving 3-dimensional latent space helps the network concurrently learn global shape regularities and local reasoning in the object coordinate space and, as a result, boosts performance.
**********

Reinforcement Learning for Control with Multiple Frequencies
Jongmin Lee, Byung-Jun Lee, Kee-Eung Kim
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Complex Dynamics in Simple Neural Networks: Understanding Gradient Flow in Phase Retrieval
Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, Lenka Zdeborová
Despite the widespread use of gradient-based algorithms for optimising high-dimensional non-convex functions, understanding their ability of finding good minima instead of being trapped in spurious ones remains to a large extent an open problem. Here we focus on gradient flow dynamics for phase retrieval from random measurements. When the ratio of the number of measurements over the input dimension is small the dynamics remains trapped in spurious minima with large basins of attraction. We find analytically that above a critical ratio those critical points become unstable developing a negative direction toward the signal. By numeric

al experiments we show that in this regime the gradient flow algorithm is not tr
apped; it drifts away from the spurious critical points along the unstable direc
tion and succeeds in finding the global minimum. Using tools from statistical ph
ysics we characterise this phenomenon, which is related to a BBP-type transition
 in the Hessian of the spurious minima.
**********

Neural Message Passing for Multi-Relational Ordered and Recursive Hypergraphs
Naganand Yadati
Message passing neural network (MPNN) has recently emerged as a successful frame
work by achieving state-of-the-art performances on many graph-based learning tas
ks.
MPNN has also recently been extended to multi-relational graphs (each edge is la
belled), and hypergraphs (each edge can connect any number of vertices).
However, in real-world datasets involving text and knowledge, relationships are
much more complex in which hyperedges can be multi-relational, recursive, and or
dered.
Such structures present several unique challenges because it is not clear how to
 adapt MPNN to variable-sized hyperedges in them.
In this work, we first unify exisiting MPNNs on different structures into G-MPNN
 (Generalised MPNN) framework.
Motivated by real-world datasets, we then propose a novel extension of the frame
work, MPNN-R (MPNN-Recursive) to handle recursively-structured data.
Experimental results demonstrate the effectiveness of proposed G-MPNN and MPNN-R
.
**********

A Unified View of Label Shift Estimation
Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, Zachary Lipton
Requests for name changes in the electronic proceedings will be accepted with 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 Private Median Estimation under Minimal Distributional Assumptions
Christos Tzamos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Ilias Zadik
We study the fundamental task of estimating the median of an underlying distribu
tion from a finite number of samples, under pure differential privacy constraint
s. We focus on distributions satisfying the minimal assumption that they have a
positive density at a small neighborhood around the median. In particular, the d
istribution is allowed to output unbounded values and is not required to have fi
nite moments. We compute the exact, up-to-constant terms, statistical rate of es
timation for the median by providing nearly-tight upper and lower bounds. Furthe
rmore, we design a polynomial-time differentially private algorithm which provab
ly achieves the optimal performance. At a technical level, our results leverage
a Lipschitz Extension Lemma which allows us to design and analyze differentially
 private algorithms solely on appropriately defined ``typical" instances of the
samples.
**********

Breaking the Communication-Privacy-Accuracy Trilemma
Wei-Ning Chen, Peter Kairouz, Ayfer Ozgur
Two major challenges in distributed learning and estimation are 1) preserving th
e privacy of the local samples; and 2) communicating them efficiently to a centr
al server, while achieving high accuracy for the end-to-end task. While there ha
s been significant interest in addressing each of these challenges separately in
 the recent literature, treatments that simultaneously address both challenges a
re still largely missing. In this paper, we develop novel encoding and decoding
mechanisms that simultaneously achieve optimal privacy and communication efficie
ncy in various canonical settings.
**********

Audeo: Audio Generation for a Silent Performance Video
Kun Su, Xiulong Liu, Eli Shlizerman

We present a novel system that gets as an input, video frames of a musician playing the piano, and generates the music for that video. The generation of music from visual cues is a challenging problem and it is not clear whether it is an attainable goal at all. Our main aim in this work is to explore the plausibility of such a transformation and to identify cues and components able to carry the association of sounds with visual events. To achieve the transformation we built a full pipeline named 'Audeo' containing three components. We first translate the video frames of the keyboard and the musician hand movements into raw mechanical musical symbolic representation Piano-Roll (Roll) for each video frame which represents the keys pressed at each time step. We then adapt the Roll to be amenable for audio synthesis by including temporal correlations. This step turns out to be critical for meaningful audio generation. In the last step, we implement Midi synthesizers to generate realistic music. Audeo converts video to audio smoothly and clearly with only a few setup constraints. We evaluate Audeo on piano performance videos collected from Youtube and obtain that their generated music is of reasonable audio quality and can be successfully recognized with high precision by popular music identification software.
**********

Ode to an ODE
Krzysztof M. Choromanski, Jared Quincy Davis, Valerii Likhosherstov, Xingyou Song, Jean-Jacques Slotine, Jacob Varley, Honglak Lee, Adrian Weller, Vikas Sindhwani
We present a new paradigm for Neural ODE algorithms, called ODEtoODE, where time-dependent parameters of the main flow evolve according to a matrix flow on the orthogonal group O(d). This nested system of two flows, where the parameter-flow is constrained to lie on the compact manifold, provides stability and effectiveness of training and solves the gradient vanishing-explosion problem which is intrinsically related to training deep neural network architectures such as Neural ODEs. Consequently, it leads to better downstream models, as we show on the example of training reinforcement learning policies with evolution strategies, and in the supervised learning setting, by comparing with previous SOTA baselines. We provide strong convergence results for our proposed mechanism that are independent of the width of the network, supporting our empirical studies. Our results show an intriguing connection between the theory of deep neural networks and the field of matrix flows on compact manifolds.
**********

Self-Distillation Amplifies Regularization in Hilbert Space
Hossein Mobahi, Mehrdad Farajtabar, Peter Bartlett
Knowledge distillation introduced in the deep learning context is a method to transfer knowledge from one architecture to another. In particular, when the architectures are identical, this is called self-distillation. The idea is to feed in predictions of the trained model as new target values for retraining (and iterate this loop possibly a few times). It has been empirically observed that the self-distilled model often achieves higher accuracy on held out data. Why this happens, however, has been a mystery: the self-distillation dynamics does not receive any new information about the task and solely evolves by looping over training. To the best of our knowledge, there is no rigorous understanding of why this happens. This work provides the first theoretical analysis of self-distillation. We focus on fitting a nonlinear function to training data, where the model space is Hilbert space and fitting is subject to L2 regularization in this function space. We show that self-distillation iterations modify regularization by progressively limiting the number of basis functions that can be used to represent the solution. This implies (as we also verify empirically) that while a few rounds of self-distillation may reduce over-fitting, further rounds may lead to under-fitting and thus worse performance.
**********

Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators
Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, Masashi Sugiyama

Invertible neural networks based on coupling flows (CF-INNs) have various machin
e learning applications such as image synthesis and representation learning. How
ever, their desirable characteristics such as analytic invertibility come at the
 cost of restricting the functional forms. This poses a question on their repres
entation power: are CF-INNs universal approximators for invertible functions? Wi
thout a universality, there could be a well-behaved invertible transformation th
at the CF-INN can never approximate, hence it would render the model class unrel
iable. We answer this question by showing a convenient criterion: a CF-INN is un
iversal if its layers contain affine coupling and invertible linear functions as
 special cases. As its corollary, we can affirmatively resolve a previously unso
lved problem: whether normalizing flow models based on affine coupling can be un
iversal distributional approximators. In the course of proving the universality,
 we prove a general theorem to show the equivalence of the universality for cert
ain diffeomorphism classes, a theoretical insight that is of interest by itself.
**********

Community detection using fast low-cardinality semidefinite programming■
Po-Wei Wang, J. Zico Kolter
Modularity maximization has been a fundamental tool for understanding the commun
ity structure of a network, but the underlying optimization problem is nonconvex
 and NP-hard to solve. State-of-the-art algorithms like the Louvain or Leiden me
thods focus on different heuristics to help escape local optima, but they still
depend on a greedy step that moves node assignment locally and is prone to getti
ng trapped. In this paper, we propose a new class of low-cardinality algorithm t
hat generalizes the local update to maximize a semidefinite relaxation derived f
rom max-k-cut. This proposed algorithm is scalable, empirically achieves the glo
bal semidefinite optimality for small cases, and outperforms the state-of-the-ar
t algorithms in real-world datasets with little additional time cost. From the a
lgorithmic perspective, it also opens a new avenue for scaling-up semidefinite p
rogramming when the solutions are sparse instead of low-rank.
**********

Modeling Noisy Annotations for Crowd Counting
Jia Wan, Antoni Chan
The annotation noise in crowd counting is not modeled in traditional crowd count
ing algorithms based on crowd density maps. In this paper, we first model the an
notation noise using a random variable with Gaussian distribution, and derive th
e pdf of the crowd density value for each spatial location in the image. We then
 approximate the joint distribution of the density values (i.e., the distributio
n of density maps) with a full covariance multivariate Gaussian density, and der
ive a low-rank approximate for tractable implementation. We use our loss functio
n to train a crowd density map estimator and achieve state-of-the-art performanc
e on three large-scale crowd counting datasets, which confirms its effectiveness
.  Examination of the predictions of the trained model shows that it can correct
ly predict the locations of people in spite of the noisy training data, which de
monstrates the robustness of our loss function to annotation noise.
**********

An operator view of policy gradient methods
Dibya Ghosh, Marlos C. Machado, Nicolas Le Roux
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations an
d Dataset Biases
Senthil Purushwalkam, Abhinav Gupta
Self-supervised representation learning approaches have recently surpassed their
 supervised learning counterparts on downstream tasks like object detection and
image classification. Somewhat mysteriously the recent gains in performance come
 from training instance classification models, treating each image and it's augm
ented versions as samples of a single class. In this work, we first present quan

titative experiments to demystify these gains. We demonstrate that approaches  l
ike MOCO and PIRL learn occlusion-invariant representations. However, they fail
to capture viewpoint and category instance invariance which are crucial componen
ts for object recognition. Second, we demonstrate that these approaches obtain f
urther gains from access to a clean object-centric training dataset like Imagene
t. Finally, we propose an approach to leverage unstructured videos to learn repr
esentations that possess higher viewpoint invariance. Our results show that the
learned representations outperform MOCOv2 trained on the same data in terms of i
nvariances encoded and the performance on downstream image classification and se
mantic segmentation tasks.
**********

Online MAP Inference of Determinantal Point Processes
Aditya Bhaskara, Amin Karbasi, Silvio Lattanzi, Morteza Zadimoghaddam
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Video Object Segmentation with Adaptive Feature Bank and Uncertain-Region Refine
ment
Yongqing Liang, Xin Li, Navid Jafari, Jim Chen
This paper presents a new matching-based framework for semi-supervised video obj
ect segmentation (VOS). Recently, state-of-the-art VOS performance has been achi
eved by matching-based algorithms, in which feature banks are created to store f
eatures for region matching and classification. However, how to effectively orga
nize information in the continuously growing feature bank remains under-explored
, and this leads to an inefficient design of the bank.  We introduced an adaptiv
e feature bank update scheme to dynamically absorb new features and discard obso
lete features. We also designed a new confidence loss and a fine-grained segment
ation module to enhance the segmentation accuracy in uncertain regions. On publi
c benchmarks, our algorithm outperforms existing state-of-the-arts.
**********

Inferring learning rules from animal decision-making
Zoe Ashwood, Nicholas A. Roy, Ji Hyun Bak, Jonathan W. Pillow
How do animals learn? This remains an elusive question in neuroscience. Whereas
reinforcement learning often focuses on the design of algorithms that enable art
ificial agents to efficiently learn new tasks, here we develop a modeling framew
ork to directly infer the empirical learning rules that animals use to acquire n
ew behaviors. Our method efficiently infers the trial-to-trial changes in an ani
mal’s policy, and decomposes those changes into a learning component and a noise
 component. Specifically, this allows us to: (i) compare different learning rule
s and objective functions that an animal may be using to update its policy; (ii)
 estimate distinct learning rates for different parameters of an animal’s policy
; (iii) identify variations in learning across cohorts of animals; and (iv) unco
ver trial-to-trial changes that are not captured by normative learning rules. Af
ter validating our framework on simulated choice data, we applied our model to d
ata from rats and mice learning perceptual decision-making tasks. We found that
certain learning rules were far more capable of explaining trial-to-trial change
s in an animal's policy. Whereas the average contribution of the conventional RE
INFORCE learning rule to the policy update for mice learning the International B
rain Laboratory's task was just 30%, we found that adding baseline parameters al
lowed the learning rule to explain 92% of the animals' policy updates under our
model. Intriguingly, the best-fitting learning rates and baseline values indicat
e that an animal's policy update, at each trial, does not occur in the direction
 that maximizes expected reward. Understanding how an animal transitions from ch
ance-level to high-accuracy performance when learning a new task not only provid
es neuroscientists with insight into their animals, but also provides concrete e
xamples of biological learning algorithms to the machine learning community.
**********

Input-Aware Dynamic Backdoor Attack

Tuan Anh Nguyen, Anh Tran

In recent years, neural backdoor attack has been considered to be a potential security threat to deep learning systems. Such systems, while achieving the state-of-the-art performance on clean data, perform abnormally on inputs with predefined triggers. Current backdoor techniques, however, rely on uniform trigger patterns, which are easily detected and mitigated by current defense methods. In this work, we propose a novel backdoor attack technique in which the triggers vary from input to input. To achieve this goal, we implement an input-aware trigger generator driven by diversity loss. A novel cross-trigger test is applied to enforce trigger nonreusablity, making backdoor verification impossible. Experiments show that our method is efficient in various attack scenarios as well as multiple datasets. We further demonstrate that our backdoor can bypass the state of the art defense methods. An analysis with a famous neural network inspector again proves the stealthiness of the proposed attack. Our code is publicly available.

**********

How hard is to distinguish graphs with graph neural networks?

Andreas Loukas

A hallmark of graph neural networks is their ability to distinguish the isomorphism class of their inputs. This study derives hardness results for the classification variant of graph isomorphism in the message-passing model (MPNN). MPNN encompasses the majority of graph neural networks used today and is universal when nodes are given unique features. The analysis relies on the introduced measure of communication capacity. Capacity measures how much information the nodes of a network can exchange during the forward pass and depends on the depth, message-size, global state, and width of the architecture. It is shown that the capacity of MPNN needs to grow linearly with the number of nodes so that a network can distinguish trees and quadratically for general connected graphs. The derived bounds concern both worst- and average-case behavior and apply to networks with/without unique features and adaptive architecture---they are also up to two orders of magnitude tighter than those given by simpler arguments. An empirical study involving 12 graph classification tasks and 420 networks reveals strong alignment between actual performance and theoretical predictions.

**********

Minimax Regret of Switching-Constrained Online Convex Optimization: No Phase Transition

Lin Chen, Qian Yu, Hannah Lawrence, 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-authors prior to requesting a name change in the electronic proceedings.

**********

Dual Manifold Adversarial Robustness: Defense against Lp and non-Lp Adversarial Attacks

Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, Soheil Feizi

Adversarial training is a popular defense strategy against attack threat models with bounded Lp norms. However, it often degrades the model performance on normal images and more importantly, the defense does not generalize well to novel attacks. Given the success of deep generative models such as GANs and VAEs in characterizing the underlying manifold of images, we investigate whether or not the aforementioned deficiencies of adversarial training can be remedied by exploiting the underlying manifold information. To partially answer this question, we consider the scenario when the manifold information of the underlying data is available. We use a subset of ImageNet natural images where an approximate underlying manifold is learned using StyleGAN. We also construct an ``On-Manifold ImageNet'' (OM-ImageNet) dataset by projecting the ImageNet samples onto the learned manifold. For OM-ImageNet, the underlying manifold information is exact. Using OM-ImageNet, we first show that on-manifold adversarial training improves both standard accuracy and robustness to on-manifold attacks. However, since no out-of-manifold perturbations are realized, the defense can be broken by Lp adversarial attacks. We further propose Dual Manifold Adversarial Training (DMAT) where adversa

rial perturbations in both latent and image spaces are used in robustifying the model. Our DMAT improves performance on normal images, and achieves comparable robustness to the standard adversarial training against Lp attacks. In addition, we observe that models defended by DMAT achieve improved robustness against novel attacks which manipulate images by global color shifts or various types of image filtering. Interestingly, similar improvements are also achieved when the defended models are tested on (out-of-manifold) natural images. These results demonstrate the potential benefits of using manifold information in enhancing robustness of deep learning models against various types of novel adversarial attacks.
**********

Cross-Scale Internal Graph Neural Network for Image Super-Resolution
Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Chen Change Loy
Non-local self-similarity in natural images has been well studied as an effective prior in image restoration. However, for single image super-resolution (SISR), most existing deep non-local methods (e.g., non-local neural networks) only exploit similar patches within the same scale of the low-resolution (LR) input image. Consequently, the restoration is limited to using the same-scale information while neglecting potential high-resolution (HR) cues from other scales. In this paper, we explore the cross-scale patch recurrence property of a natural image, i.e., similar patches tend to recur many times across different scales. This is achieved using a novel cross-scale internal graph neural network (IGNN). Specifically, we dynamically construct a cross-scale graph by searching k-nearest neighboring patches in the downsampled LR image for each query patch in the LR image. We then obtain the corresponding k HR neighboring patches in the LR image and aggregate them adaptively in accordance to the edge label of the constructed graph. In this way, the HR information can be passed from k HR neighboring patches to the LR query patch to help it recover more detailed textures. Besides, these internal image-specific LR/HR exemplars are also significant complements to the external information learned from the training dataset. Extensive experiments demonstrate the effectiveness of IGNN against the state-of-the-art SISR methods including existing non-local networks on standard benchmarks.
**********

Unsupervised Representation Learning by Invariance Propagation
Feng Wang, Huaping Liu, Di Guo, Sun Fuchun
Unsupervised learning methods based on contrastive learning have drawn increasing attention and achieved promising results. Most of them aim to learn representations invariant to instance-level variations, which are provided by different views of the same instance. In this paper, we propose Invariance Propagation to focus on learning representations invariant to category-level variations, which are provided by different instances from the same category. Our method recursively discovers semantically consistent samples residing in the same high-density regions in representation space. We demonstrate a hard sampling strategy to concentrate on maximizing the agreement between the anchor sample and its hard positive samples, which provide more intra-class variations to help capture more abstract invariance. As a result, with a ResNet-50 as the backbone, our method achieves 71.3% top-1 accuracy on ImageNet linear classification and 78.2% top-5 accuracy fine-tuning on only 1% labels, surpassing previous results. We also achieve state-of-the-art performance on other downstream tasks, including linear classification on Places205 and Pascal VOC, and transfer learning on small scale datasets.

**********

Restoring Negative Information in Few-Shot Object Detection
Yukuan Yang, Fangyun Wei, Miaojing Shi, Guoqi Li
Few-shot learning has recently emerged as a new challenge in the deep learning field: unlike conventional methods that train the deep neural networks (DNNs) with a large number of labeled data, it asks for the generalization of DNNs on new classes with few annotated samples. Recent advances in few-shot learning mainly focus on image classification while in this paper we focus on object detection. The initial explorations in few-shot object detection tend to simulate a classification scenario by using the positive proposals in images with respect to certa

in object class while discarding the negative proposals of that class. Negatives , especially hard negatives, however, are essential to the embedding space learn ing in few-shot object detection. In this paper, we restore the negative informa tion in few-shot object detection by introducing a new negative- and positive-re presentative based metric learning framework and a new inference scheme with neg ative and positive representatives. We build our work on a recent few-shot pipel ine RepMet with several new modules to encode negative information for both trai ning and testing. Extensive experiments on ImageNet-LOC and PASCAL VOC show our method substantially improves the state-of-the-art few-shot object detection sol utions. Our code is available at https://github.com/yang-yk/NP-RepMet.
**********

Do Adversarially Robust ImageNet Models Transfer Better?
Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, Aleksander Madry
Transfer learning is a widely-used paradigm in deep learning, where models pre-t rained on standard datasets can be efficiently adapted to downstream tasks. Typi cally, better pre-trained models yield better transfer results, suggesting that initial accuracy is a key aspect of transfer learning performance. In this work, we identify another such aspect: we find that adversarially robust models, whil e less accurate, often perform better than their standard-trained counterparts w hen used for transfer learning. Specifically, we focus on adversarially robust I mageNet classifiers, and show that they yield improved accuracy on a standard su ite of downstream classification tasks. Further analysis uncovers more differenc es between robust and standard models in the context of transfer learning. Our r esults are consistent with (and in fact, add to) recent hypotheses stating that robustness leads to improved feature representations. Code and models is availab le in the supplementary material.
**********

Robust Correction of Sampling Bias using Cumulative Distribution Functions
Bijan Mazaheri, Siddharth Jain, Jehoshua Bruck
Varying domains and biased datasets can lead to differences between the training and the target distributions, known as covariate shift. Current approaches for alleviating this often rely on estimating the ratio of training and target proba bility density functions. These techniques require parameter tuning and can be u nstable across different datasets. We present a new method for handling covariat e shift using the empirical cumulative distribution function estimates of the ta rget distribution by a rigorous generalization of a recent idea proposed by Vapn ik and Izmailov. Further, we show experimentally that our method is more robust in its predictions, is not reliant on parameter tuning and shows similar classif ication performance compared to the current state-of-the-art techniques on synth etic and real datasets.
**********

Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Me ta-Learning Approach
Alireza Fallah, Aryan Mokhtari, Asuman Ozdaglar
In Federated Learning, we aim to train models across multiple computing units (u sers), while users can only communicate with a common central server, without ex changing their data samples. This mechanism exploits the computational power of all users and allows users to obtain a richer model as their models are trained over a larger set of data points. However, this scheme only develops a common ou tput for all the users, and, therefore, it does not adapt the model to each user . This is an important missing feature, especially given the heterogeneity of th e underlying data distribution for various users. In this paper, we study a pers onalized variant of the federated learning in which our goal is to find an initi al shared model that current or new users can easily adapt to their local datase t by performing one or a few steps of gradient descent with respect to their own data. This approach keeps all the benefits of the federated learning architectu re, and, by structure, leads to a more personalized model for each user. We show this problem can be studied within the Model-Agnostic Meta-Learning (MAML) fram ework. Inspired by this connection, we study a personalized variant of the well-known Federated Averaging algorithm and evaluate its performance in terms of gra

dient norm for non-convex loss functions. Further, we characterize how this perf
ormance is affected by the closeness of underlying distributions of user data, m
easured in terms of distribution distances such as Total Variation and 1-Wassers
tein metric.
**********

Pixel-Level Cycle Association: A New Perspective for Domain Adaptive Semantic Se
gmentation
Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, Alexander Hauptmann
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Classification with Valid and Adaptive Coverage
Yaniv Romano, Matteo Sesia, Emmanuel Candes
Conformal inference, cross-validation+, and the jackknife+ are hold-out methods
that can be combined with virtually any machine learning algorithm to construct
prediction sets with guaranteed marginal coverage. In this paper, we develop spe
cialized versions of these techniques for categorical and unordered response lab
els that, in addition to providing marginal coverage, are also fully adaptive to
 complex data distributions, in the sense that they perform favorably in terms o
f approximate conditional coverage compared to alternative methods. The heart of
 our contribution is a novel conformity score, which we explicitly demonstrate t
o be powerful and intuitive for classification problems, but whose underlying pr
inciple is potentially far more general. Experiments on synthetic and real data
demonstrate the practical value of our theoretical guarantees, as well as the st
atistical advantages of the proposed methods over the existing alternatives.
**********

Learning Global Transparent Models consistent with Local Contrastive Explanation
s
Tejaswini Pedapati, Avinash Balakrishnan, Karthikeyan Shanmugam, Amit Dhurandhar
There is a rich and growing literature on producing local contrastive/counterfac
tual explanations for black-box models (e.g. neural networks). In these methods,
 for an input, an explanation is in the form of a contrast point differing in ve
ry few features from the original input and lying in a different class. Other wo
rks try to build globally interpretable models like decision trees and rule list
s based on the data using actual labels or based on the black-box models predict
ions. Although these interpretable global models can be useful, they may not be
consistent with local explanations from a specific black-box of choice. In this
work, we explore the question: Can we produce a transparent global model that is
 simultaneously accurate and consistent with the local (contrastive) explanation
s of the black-box model? We introduce a local consistency metric that quantifie
s if the local explanations for the black-box model are also applicable to the p
roxy/surrogate globally transparent model. Based on a key insight we propose a n
ovel method where we create custom boolean features from local contrastive expla
nations of the black-box model and then train a globally transparent model that
has higher local consistency compared with other known strategies in addition to
 being accurate.
**********

Learning to Approximate a Bregman Divergence
Ali Siahkamari, XIDE XIA, Venkatesh Saligrama, David Castañón, Brian Kulis
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Diverse Image Captioning with Context-Object Split Latent Spaces
Shweta Mahajan, Stefan Roth
Diverse image captioning models aim to learn one-to-many mappings that are innat
e to cross-domain datasets, such as of images and texts. Current methods for thi

s task are based on  generative latent variable models, eg. VAEs with structured  latent spaces. Yet, the amount of multimodality captured by prior work is limit ed to that of the paired training data -- the true diversity of the underlying g enerative process is not fully captured. To address this limitation, we leverage  the contextual descriptions in the dataset that explain similar contexts in dif ferent visual scenes. To this end, we introduce a novel factorization of the lat ent space, termed context-object split, to model diversity in contextual descrip tions across images and texts within the dataset. Our framework not only enables  diverse captioning through context-based pseudo supervision, but extends this t o images with novel objects and without paired captions in the training data. We  evaluate our COS-CVAE approach on the standard COCO dataset and on the held-out  COCO dataset consisting of images with novel objects, showing significant gains  in accuracy and diversity.
**********

Learning Disentangled Representations of Videos with Missing Data
Armand Comas, Chi Zhang, Zlatan Feric, Octavia Camps, Rose Yu
Missing data poses significant challenges while learning representations of vide o sequences. We present Disentangled Imputed Video autoEncoder (DIVE), a deep ge nerative model that imputes and predicts future video frames in the presence of  missing data. Specifically, DIVE introduces a missingness latent variable,  dise ntangles the hidden video representations into static and dynamic appearance, po se, and missingness factors for each object, while it imputes each object trajec tory where data is missing.
On a moving MNIST  dataset with various missing scenarios,  DIVE outperforms the  state of the art baselines by a substantial margin.  We also present comparison s on a real-world MOTSChallenge pedestrian dataset, which demonstrates the pract ical value of our method in a more realistic setting. Our code can be found in h ttps://github.com/Rose-STL-Lab/DIVE.
**********

Natural Graph Networks
Pim de Haan, Taco S. Cohen, Max Welling
A key requirement for graph neural networks is that they must process a graph in  a way that does not depend on how the graph is described. Traditionally this ha s been taken to mean that a graph network must be equivariant to node permutatio ns. Here we show that instead of equivariance, the more general concept of natur ality is sufficient for a graph network to be well-defined, opening up a larger  class of graph networks. We define global and local natural graph networks, the  latter of which are as scalable as conventional message passing graph neural net works while being more flexible. We give one practical instantiation of a natura l network on graphs which uses an equivariant message network parameterization,  yielding good performance on several benchmarks.
**********

Continual Learning with Node-Importance based Adaptive Group Sparse Regularizati on
Sangwon Jung, Hongjoon Ahn, Sungmin Cha, Taesup Moon
We propose a novel regularization-based continual learning method, dubbed as Ada ptive Group Sparsity based Continual Learning (AGS-CL), using two group sparsity -based penalties. Our method selectively employs the two penalties when learning  each neural network node based on its the importance, which is adaptively updat ed after learning each task. By utilizing the proximal gradient descent method,  the exact sparsity and freezing of the model is guaranteed during the learning p rocess, and thus, the learner explicitly controls the model capacity. Furthermor e, as a critical detail, we re-initialize the weights associated with unimportan t nodes after learning each task in order to facilitate efficient learning and p revent the negative transfer. Throughout the extensive experimental results, we  show that our AGS-CL uses orders of magnitude less memory space for storing the  regularization parameters, and it significantly outperforms several state-of-the -art baselines on representative benchmarks for both supervised and reinforcemen t learning.
**********

Towards Crowdsourced Training of Large Neural Networks using Decentralized Mixtu
re-of-Experts
Max Ryabinin, Anton Gusev
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Bidirectional Convolutional Poisson Gamma Dynamical Systems
wenchao chen, Chaojie Wang, Bo Chen, Yicheng Liu, Hao Zhang, Mingyuan Zhou
Incorporating the natural document-sentence-word structure into hierarchical Bay
esian modeling, we propose convolutional Poisson gamma dynamical systems (PGDS)
that introduce not only word-level probabilistic convolutions, but also sentence
-level stochastic temporal transitions. With word-level convolutions capturing p
hrase-level topics and sentence-level transitions capturing how the topic usages
 evolve over consecutive sentences, we aggregate the topic proportions of all se
ntences of a document as its feature representation. To consider not only forwar
d but also backward sentence-level information transmissions, we further develop
 a bidirectional convolutional PGDS to incorporate the full contextual informati
on to represent each sentence. For efficient inference, we construct a convoluti
onal-recurrent inference network, which provides both sentence-level and documen
t-level representations, and introduce a hybrid Bayesian inference scheme combin
ing stochastic-gradient MCMC and amortized variational inference. Experimental r
esults on a variety of document corpora demonstrate that the proposed models can
 extract expressive multi-level latent representations, including interpretable
phrase-level topics and sentence-level temporal transitions as well as discrimin
ative document-level features, achieving state-of-the-art document categorizatio
n performance while being memory and computation efficient.
**********
Deep Reinforcement and InfoMax Learning
Bogdan Mazoure, Remi Tachet des Combes, Thang Long Doan, Philip Bachman, R Devon
 Hjelm
We posit that a reinforcement learning (RL) agent will perform better when it us
es representations that are better at predicting the future, particularly in ter
ms of few-shot learning and domain adaptation. To test that hypothesis, we intro
duce an objective based on Deep InfoMax (DIM) which trains the agent to predict
the future by maximizing the mutual information between its internal representat
ion of successive timesteps. We provide an intuitive analysis of the convergence
 properties of our approach from the perspective of Markov chain mixing times, a
nd argue that convergence of the lower bound on mutual information is related to
 the inverse absolute spectral gap of the transition model. We test our approach
 in several synthetic settings, where it successfully learns representations tha
t are predictive of the future. Finally, we augment C51, a strong distributional
 RL agent, with our temporal DIM objective and demonstrate on a continual learni
ng task (inspired by Ms.~PacMan) and on the recently introduced Procgen environm
ent that our approach improves performance, which supports our core hypothesis.
**********
On ranking via sorting by estimated expected utility
Clement Calauzenes, Nicolas Usunier
Ranking and selection tasks appear in different contexts with specific desiderat
a, such as the maximizaton of average relevance on the top of the list, the requ
irement of diverse rankings, or, relatedly, the focus on providing at least one
relevant items to as many users as possible. This paper addresses the question o
f which of these tasks are asymptotically solved by sorting by decreasing order
of expected utility, for some suitable notion of utility, or, equivalently, \emp
h{when is square loss regression consistent for ranking \emph{via} score-and-sor
t?}. We provide an answer to this question in the form of a structural character
ization of ranking losses for which a suitable regression is consistent. This re
sult has two fundamental corollaries. First, whenever there exists a consistent
approach based on convex risk minimization, there also is a consistent approach

based on regression. Second, when regression is not consistent, there are data distributions for which consistent surrogate approaches necessarily have non-trivial local minima, and optimal scoring function are necessarily discontinuous, even when the underlying data distribution is regular. In addition to providing a better understanding of surrogate approaches for ranking, these results illustrate the intrinsic difficulty of solving general ranking problems with the score-and-sort approach.
**********
Distribution-free binary classification: prediction sets, confidence intervals and calibration
Chirag Gupta, Aleksandr Podkopaev, Aaditya Ramdas
We study three notions of uncertainty quantification---calibration, confidence intervals and prediction sets---for binary classification in the distribution-free setting, that is without making any distributional assumptions on the data. With a focus towards calibration, we establish a 'tripod' of theorems that connect these three notions for score-based classifiers. A direct implication is that distribution-free calibration is only possible, even asymptotically, using a scoring function whose level sets partition the feature space into at most countably many sets. Parametric calibration schemes such as variants of Platt scaling do not satisfy this requirement, while nonparametric schemes based on binning do. To close the loop, we derive distribution-free confidence intervals for binned probabilities for both fixed-width and uniform-mass binning. As a consequence of our 'tripod' theorems, these confidence intervals for binned probabilities lead to distribution-free calibration. We also derive extensions to settings with streaming data and covariate shift.
**********
Closing the Dequantization Gap: PixelCNN as a Single-Layer Flow
Didrik Nielsen, Ole Winther
Flow models have recently made great progress at modeling ordinal discrete data such as images and audio. Due to the continuous nature of flow models, dequantization is typically applied when using them for such discrete data, resulting in lower bound estimates of the likelihood. In this paper, we introduce subset flows, a class of flows that can tractably transform finite volumes and thus allow exact computation of likelihoods for discrete data. Based on subset flows, we identify ordinal discrete autoregressive models, including WaveNets, PixelCNNs and Transformers, as single-layer flows. We use the flow formulation to compare models trained and evaluated with either the exact likelihood or its dequantization lower bound. Finally, we study multilayer flows composed of PixelCNNs and non-autoregressive coupling layers and demonstrate state-of-the-art results on CIFAR-10 for flow models trained with dequantization.
**********
Sequence to Multi-Sequence Learning via Conditional Chain Mapping for Mixture Signals
Jing Shi, Xuankai Chang, Pengcheng Guo, Shinji Watanabe, Yusuke Fujita, Jiaming Xu, Bo Xu, Lei Xie
Neural sequence-to-sequence models are well established for applications which can be cast as mapping a single input sequence into a single output sequence. In this work, we focus on one-to-many sequence transduction problems, such as extracting multiple sequential sources from a mixture sequence. We extend the standard sequence-to-sequence model to a conditional multi-sequence model, which explicitly models the relevance between multiple output sequences with the probabilistic chain rule. Based on this extension, our model can conditionally infer output sequences one-by-one by making use of both input and previously-estimated contextual output sequences. This model additionally has a simple and efficient stop criterion for the end of the transduction, making it able to infer the variable number of output sequences. We take speech data as a primary test field to evaluate our methods since the observed speech data is often composed of multiple sources due to the nature of the superposition principle of sound waves. Experiments on several different tasks including speech separation and multi-speaker speech recognition show that our conditional multi-sequence models lead to consisten

t improvements over the conventional non-conditional models.
**********
Variance reduction for Random Coordinate Descent-Langevin Monte Carlo
ZHIYAN DING, Qin Li
Sampling from a log-concave distribution function is one core problem that has wide applications in Bayesian statistics and machine learning. While most gradient free methods have slow convergence rate, the Langevin Monte Carlo (LMC) that provides fast convergence requires the computation of gradients. In practice one uses finite-differencing approximations as surrogates, and the method is expensive in high-dimensions.
**********
Language as a Cognitive Tool to Imagine Goals in Curiosity Driven Exploration
Cédric Colas, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter Dominey, Pierre-Yves Oudeyer
Developmental machine learning studies how artificial agents can model the way children learn open-ended repertoires of skills. Such agents need to create and represent goals, select which ones to pursue and learn to achieve them. Recent approaches have considered goal spaces that were either fixed and hand-defined or learned using generative models of states. This limited agents to sample goals within the distribution of known effects. We argue that the ability to imagine out-of-distribution goals is key to enable creative discoveries and open-ended learning. Children do so by leveraging the compositionality of language as a tool to imagine descriptions of outcomes they never experienced before, targeting them as goals during play. We introduce IMAGINE, an intrinsically motivated deep reinforcement learning architecture that models this ability. Such imaginative agents, like children, benefit from the guidance of a social peer who provides language descriptions. To take advantage of goal imagination, agents must be able to leverage these descriptions to interpret their imagined out-of-distribution goals. This generalization is made possible by modularity: a decomposition between learned goal-achievement reward function and policy relying on deep sets, gated attention and object-centered representations. We introduce the Playground environment and study how this form of goal imagination improves generalization and exploration over agents lacking this capacity. In addition, we identify the properties of goal imagination that enable these results and study the impacts of modularity and social interactions.
**********
All Word Embeddings from One Embedding
Sho Takase, Sosuke Kobayashi
In neural network-based models for natural language processing (NLP), the largest part of the parameters often consists of word embeddings. Conventional models prepare a large embedding matrix whose size depends on the vocabulary size. Therefore, storing these models in memory and disk storage is costly. In this study, to reduce the total number of parameters, the embeddings for all words are represented by transforming a shared embedding. The proposed method, ALONE (all word embeddings from one), constructs the embedding of a word by modifying the shared embedding with a filter vector, which is word-specific but non-trainable. Then, we input the constructed embedding into a feed-forward neural network to increase its expressiveness. Naively, the filter vectors occupy the same memory size as the conventional embedding matrix, which depends on the vocabulary size. To solve this issue, we also introduce a memory-efficient filter construction approach. We indicate our ALONE can be used as word representation sufficiently through an experiment on the reconstruction of pre-trained word embeddings. In addition, we also conduct experiments on NLP application tasks: machine translation and summarization. We combined ALONE with the current state-of-the-art encoder-decoder model, the Transformer [36], and achieved comparable scores on WMT 2014 English-to-German translation and DUC 2004 very short summarization with less parameters.
**********
Primal Dual Interpretation of the Proximal Stochastic Gradient Langevin Algorith

m
Adil Salim, Peter Richtarik
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

How to Characterize The Landscape of Overparameterized Convolutional Neural Networks
Yihong Gu, Weizhong Zhang, Cong Fang, Jason D. Lee, Tong Zhang
For many initialization schemes, parameters of two randomly initialized deep neural networks (DNNs) can be quite different, but feature distributions of the hidden nodes are similar at each layer. With the help of a new technique called {\it neural network grafting}, we demonstrate that even during the entire training process, feature distributions of differently initialized networks remain similar at each layer. In this paper, we present an explanation of this phenomenon. Specifically, we consider the loss landscape of an overparameterized convolutional neural network (CNN) in the continuous limit, where the numbers of channels/hidden nodes in the hidden layers go to infinity. Although the landscape of the overparameterized CNN is still non-convex with respect to the trainable parameters, we show that very surprisingly, it can be reformulated as a convex function with respect to the feature distributions in the hidden layers. Therefore by reparameterizing neural networks in terms of feature distributions, we obtain a much simpler characterization of the landscape of overparameterized CNNs. We further argue that training with respect to network parameters leads to a fixed trajectory in the feature distributions.
**********

On the Tightness of Semidefinite Relaxations for Certifying Robustness to Adversarial Examples
Richard Zhang
The robustness of a neural network to adversarial examples can be provably certified by solving a convex relaxation. If the relaxation is loose, however, then the resulting certificate can be too conservative to be practically useful. Recently, a less conservative robustness certificate was proposed, based on a semidefinite programming (SDP) relaxation of the ReLU activation function. In this paper, we describe a geometric technique that determines whether this SDP certificate is exact, meaning whether it provides both a lower-bound on the size of the smallest adversarial perturbation, as well as a globally optimal perturbation that attains the lower-bound. Concretely, we show, for a least-squares restriction of the usual adversarial attack problem, that the SDP relaxation amounts to the nonconvex projection of a point onto a hyperbola. The resulting SDP certificate is exact if and only if the projection of the point lies on the major axis of the hyperbola. Using this geometric technique, we prove that the certificate is exact over a single hidden layer under mild assumptions, and explain why it is usually conservative for several hidden layers. We experimentally confirm our theoretical insights using a general-purpose interior-point method and a custom rank-2 Burer-Monteiro algorithm.
**********

Submodular Meta-Learning
Arman Adibi, Aryan Mokhtari, Hamed Hassani
In this paper, we introduce a discrete variant of the Meta-learning framework. Meta-learning aims at exploiting prior experience and data to improve performance on future tasks. By now, there exist numerous formulations for Meta-learning in the continuous domain. Notably, the Model-Agnostic Meta-Learning (MAML) formulation views each task as a continuous optimization problem and based on prior data learns a suitable initialization that can be adapted to new, unseen tasks after a few simple gradient updates. Motivated by this terminology, we propose a novel Meta-learning framework in the discrete domain where each task is equivalent to maximizing a set function under a cardinality constraint. Our approach aims at using prior data, i.e., previously visited tasks, to train a proper initial so

lution set that can be quickly adapted to a new task at a relatively low computa tional cost. This approach leads to (i) a personalized solution for each task, a nd (ii) significantly reduced computational cost at test time compared to the ca se where the solution is fully optimized once the new task is revealed. The trai ning procedure is performed by solving a challenging discrete optimization probl em for which we present deterministic and randomized algorithms. In the case whe re the tasks are monotone and submodular, we show strong theoretical guarantees for our proposed methods even though the training objective may not be submodula r. We also demonstrate the effectiveness of our framework on two real-world prob lem instances where we observe that our methods lead to a significant reduction in computational complexity in solving the new tasks while incurring a small per formance loss compared to when the tasks are fully optimized.

**********

Rethinking Pre-training and Self-training

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk , Quoc Le

Pre-training is a dominant paradigm in computer vision. For example, supervised ImageNet pre-training is commonly used to initialize the backbones of object det ection and segmentation models. He et al., however, show a striking result that ImageNet pre-training has limited impact on COCO object detection. Here we inves tigate self-training as another method to utilize additional data on the same se tup and contrast it against ImageNet pre-training. Our study reveals the general ity and flexibility of self-training with three additional insights: 1) stronger data augmentation and more labeled data further diminish the value of pre-train ing, 2) unlike pre-training, self-training is always helpful when using stronger data augmentation, in both low-data and high-data regimes, and 3) in the case t hat pre-training is helpful, self-training improves upon pre-training. For examp le, on the COCO object detection dataset, pre-training benefits when we use one fifth of the labeled data, and hurts accuracy when we use all labeled data. Self -training, on the other hand, shows positive improvements from +1.3 to +3.4AP ac ross all dataset sizes. In other words, self-training works well exactly on the same setup that pre-training does not work (using ImageNet to help COCO). On the PASCAL segmentation dataset, which is a much smaller dataset than COCO, though pre-training does help significantly, self-training improves upon the pre-traine d model. On COCO object detection, we achieve 53.8AP, an improvement of +1.7AP o ver the strongest SpineNet model. On PASCAL segmentation, we achieve 90.5mIOU, a n improvement of +1.5mIOU over the previous state-of-the-art result by DeepLabv3 +.

**********

Unsupervised Sound Separation Using Mixture Invariant Training

Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, John Her shey

In recent years, rapid progress has been made on the problem of single-channel s ound separation using supervised training of deep neural networks. In such super vised approaches, a model is trained to predict the component sources from synth etic mixtures created by adding up isolated ground-truth sources. Reliance on th is synthetic training data is problematic because good performance depends upon the degree of match between the training data and real-world audio, especially i n terms of the acoustic conditions and distribution of sources. The acoustic pro perties can be challenging to accurately simulate, and the distribution of sound types may be hard to replicate. In this paper, we propose a completely unsuperv ised method, mixture invariant training (MixIT), that requires only single-chann el acoustic mixtures. In MixIT, training examples are constructed by mixing toge ther existing mixtures, and the model separates them into a variable number of l atent sources, such that the separated sources can be remixed to approximate the original mixtures. We show that MixIT can achieve competitive performance compa red to supervised methods on speech separation. Using MixIT in a semi-supervised learning setting enables unsupervised domain adaptation and learning from large amounts of real-world data without ground-truth source waveforms. In particular , we significantly improve reverberant speech separation performance by incorpor

ating reverberant mixtures, train a speech enhancement system from noisy mixtures, and improve universal sound separation by incorporating a large amount of in-the-wild data.
**********

Adaptive Discretization for Model-Based Reinforcement Learning
Sean Sinclair, Tianyu Wang, Gauri Jain, Siddhartha Banerjee, Christina Yu
We introduce the technique of adaptive discretization to design an efficient model-based episodic reinforcement learning algorithm in large (potentially continuous) state-action spaces. Our algorithm is based on optimistic one-step value iteration extended to maintain an adaptive discretization of the space. From a theoretical perspective we provide worst-case regret bounds for our algorithm which are competitive compared to the state-of-the-art model-based algorithms. Moreover, our bounds are obtained via a modular proof technique which can potentially extend to incorporate additional structure on the problem.
**********

CodeCMR: Cross-Modal Retrieval For Function-Level Binary Source Code Matching
Zeping Yu, Wenxin Zheng, Jiaqi Wang, Qiyi Tang, Sen Nie, Shi Wu
Binary source code matching, especially on function-level, has a critical role in the field of computer security. Given binary code only, finding the corresponding source code improves the accuracy and efficiency in reverse engineering. Given source code only, related binary code retrieval contributes to known vulnerabilities confirmation. However, due to the vast difference between source and binary code, few studies have investigated binary source code matching. Previously published studies focus on code literals extraction such as strings and integers, then utilize traditional matching algorithms such as the Hungarian algorithm for code matching. Nevertheless, these methods have limitations on function-level, because they ignore the potential semantic features of code and a lot of code lacks sufficient code literals. Also, these methods indicate a need for expert experience for useful feature identification and feature engineering, which is timeconsuming. This paper proposes an end-to-end cross-modal retrieval network for binary source code matching, which achieves higher accuracy and requires less expert experience. We adopt Deep Pyramid Convolutional Neural Network (DPCNN) for source code feature extraction and Graph Neural Network (GNN) for binary code feature extraction. We also exploit neural network-based models to capture code literals, including strings and integers. Furthermore, we implement "norm weighted sampling" for negative sampling. We evaluate our model on two datasets, where it outperforms other methods significantly.
**********

On Warm-Starting Neural Network Training
Jordan Ash, Ryan P. Adams
In many real-world deployments of machine learning systems, data arrive piecemeal. These learning scenarios may be passive, where data arrive incrementally due to structural properties of the problem (e.g., daily financial data) or active, where samples are selected according to a measure of their quality (e.g., experimental design). In both of these cases, we are building a sequence of models that incorporate an increasing amount of data. We would like each of these models in the sequence to be performant and take advantage of all the data that are available to that point. Conventional intuition suggests that when solving a sequence of related optimization problems of this form, it should be possible to initialize using the solution of the previous iterate---to ``warm start'' the optimization rather than initialize from scratch---and see reductions in wall-clock time. However, in practice this warm-starting seems to yield poorer generalization performance than models that have fresh random initializations, even though the final training losses are similar. While it appears that some hyperparameter settings allow a practitioner to close this generalization gap, they seem to only do so in regimes that damage the wall-clock gains of the warm start. Nevertheless, it is highly desirable to be able to warm-start neural network training, as it would dramatically reduce the resource usage associated with the construction of performant deep learning systems. In this work, we take a closer look at this empirical phenomenon and try to understand when and how it occurs. We also provid

e a surprisingly simple trick that overcomes this pathology in several important situations, and present experiments that elucidate some of its properties.
**********
DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks
Dennis Wei, Tian Gao, Yue Yu
This paper re-examines a continuous optimization framework dubbed NOTEARS for learning Bayesian networks. We first generalize existing algebraic characterizations of acyclicity to a class of matrix polynomials. Next, focusing on a one-parameter-per-edge setting, it is shown that the Karush-Kuhn-Tucker (KKT) optimality conditions for the NOTEARS formulation cannot be satisfied except in a trivial case, which explains a behavior of the associated algorithm. We then derive the KKT conditions for an equivalent reformulation, show that they are indeed necessary, and relate them to explicit constraints that certain edges be absent from the graph. If the score function is convex, these KKT conditions are also sufficient for local minimality despite the non-convexity of the constraint. Informed by the KKT conditions, a local search post-processing algorithm is proposed and shown to substantially and universally improve the structural Hamming distance of all tested algorithms, typically by a factor of 2 or more. Some combinations with local search are both more accurate and more efficient than the original NOTEARS.
**********
OOD-MAML: Meta-Learning for Few-Shot Out-of-Distribution Detection and Classification
Taewon Jeong, Heeyoung Kim
We propose a few-shot learning method for detecting out-of-distribution (OOD) samples from classes that are unseen during training while classifying samples from seen classes using only a few labeled examples. For detecting unseen classes while generalizing to new samples of known classes, we synthesize fake samples, i.e., OOD samples, but that resemble in-distribution samples, and use them along with real samples. Our approach is based on an extension of model-agnostic meta learning (MAML) and is denoted as OOD-MAML, which not only learns a model initialization but also the initial fake samples across tasks. The learned initial fake samples can be used to quickly adapt to new tasks to form task-specific fake samples with only one or a few gradient update steps using MAML. For testing, OOD-MAML converts a K-shot N-way classification
task into N sub-tasks of K-shot OOD detection with respect to each class. The joint analysis of N sub-tasks facilitates simultaneous classification and OOD detection and, furthermore, offers an advantage, in that it does not require re-training when the number of classes for a test task differs from that for training tasks; it is sufficient to simply assume as many sub-tasks as the number of classes for the test task. We also demonstrate the effective performance of OOD-MAML over benchmark datasets.
**********
An Imitation from Observation Approach to Transfer Learning with Dynamics Mismatch
Siddharth Desai, Ishan Durugkar, Haresh Karnan, Garrett Warnell, Josiah Hanna, Peter Stone
We examine the problem of transferring a policy learned in a source environment to a target environment with different dynamics, particularly in the case where it is critical to reduce the amount of interaction with the target environment during learning. This problem is particularly important in sim-to-real transfer because simulators inevitably model real-world dynamics imperfectly. In this paper, we show that one existing solution to this transfer problem-- grounded action transformation --is closely related to the problem of imitation from observation (IfO): learning behaviors that mimic the observations of behavior demonstrations. After establishing this relationship, we hypothesize that recent state-of-the-art approaches from the IfO literature can be effectively repurposed for grounded transfer learning. To validate our hypothesis we derive a new algorithm -- generative adversarial reinforced action transformation (GARAT) -- based on adver

sarial imitation from observation techniques. We run experiments in several doma ins with mismatched dynamics, and find that agents trained with GARAT achieve hi gher returns in the target environment compared to existing black-box transfer m ethods.
**********

Learning About Objects by Learning to Interact with Them
Martin Lohmann, Jordi Salvador, Aniruddha Kembhavi, Roozbeh Mottaghi
Much of the remarkable progress in computer vision has been focused around fully supervised learning mechanisms relying on highly curated datasets for a variety of tasks. In contrast, humans often learn about their world with little to no e xternal supervision. Taking inspiration from infants learning from their environ ment through play and interaction, we present a computational framework to disco ver objects and learn their physical properties along this paradigm of Learning from Interaction. Our agent, when placed within the near photo-realistic and phy sics-enabled AI2-THOR environment, interacts with its world and learns about obj ects, their geometric extents and relative masses, without any external guidance . Our experiments reveal that this agent learns efficiently and effectively; not just for objects it has interacted with before, but also for novel instances fr om seen categories as well as novel object categories.
**********

Learning discrete distributions with infinite support
Doron Cohen, Aryeh Kontorovich, Geo■rey Wolfer
We present a novel approach to estimating discrete distributions with (potential ly) infinite support in the total variation metric. In a departure from the esta blished paradigm, we make no structural assumptions whatsoever on the sampling d istribution. In such a setting, distribution-free risk bounds are impossible, an d the best one could hope for is a fully empirical data-dependent bound. We deri ve precisely such bounds, and demonstrate that these are, in a well-defined sens e, the best possible. Our main discovery is that the half-norm of the empirical distribution provides tight upper and lower estimates on the empirical risk. Fur thermore, this quantity decays at a nearly optimal rate as a function of the tru e distribution. The optimality follows from a minimax result, of possible indepe ndent interest. Additional structural results are provided, including an exact R ademacher complexity calculation and apparently a first connection between the t otal variation risk and the missing mass.
**********

Dissecting Neural ODEs
Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, Hajime Asama
Continuous deep learning architectures have recently re-emerged as Neural Ordina ry Differential Equations (Neural ODEs). This infinite-depth approach theoretica lly bridges the gap between deep learning and dynamical systems, offering a nove l perspective. However, deciphering the inner working of these models is still a n open challenge, as most applications apply them as generic black-box modules. In this work we ``open the box'', further developing the continuous-depth formul ation with the aim of clarifying the influence of several design choices on the underlying dynamics.
**********

Teaching a GAN What Not to Learn
Siddarth Asokan, Chandra Seelamantula
Generative adversarial networks (GANs) were originally envisioned as unsupervise d generative models that learn to follow a target distribution. Variants such as conditional GANs, auxiliary-classifier GANs (ACGANs) project GANs on to supervi sed and semi-supervised learning frameworks by providing labelled data and using multi-class discriminators. In this paper, we approach the supervised GAN probl em from a different perspective, one that is motivated by the philosophy of the famous Persian poet Rumi who said, "The art of knowing is knowing what to ignore ." In the GAN framework, we not only provide the GAN positive data that it must learn to model, but also present it with so-called negative samples that it must learn to avoid — we call this "The Rumi Framework." This formulation allows the discriminator to represent the underlying target distribution better by learnin

g to penalize generated samples that are undesirable – we show that this capabil
ity accelerates the learning process of the generator. We present a reformulatio
n of the standard GAN (SGAN) and least-squares GAN (LSGAN) within the Rumi setti
ng. The advantage of the reformulation is demonstrated by means of experiments c
onducted on MNIST, Fashion MNIST, CelebA, and CIFAR-10 datasets. Finally, we con
sider an application of the proposed formulation to address the important proble
m of learning an under-represented class in an unbalanced dataset. The Rumi appr
oach results in substantially lower FID scores than the standard GAN frameworks
while possessing better generalization capability.
**********

Counterfactual Data Augmentation using Locally Factored Dynamics
Silviu Pitis, Elliot Creager, Animesh Garg
Many dynamic processes, including common scenarios in robotic control and reinfo
rcement learning (RL), involve a set of interacting subprocesses. Though the sub
processes are not independent, their interactions are often sparse, and the dyna
mics at any given time step can often be decomposed into locally independent} ca
usal mechanisms. Such local causal structures can be leveraged to improve the sa
mple efficiency of sequence prediction and off-policy reinforcement learning. We
 formalize this by introducing local causal models (LCMs), which are induced fro
m a global causal model by conditioning on a subset of the state space. We propo
se an approach to inferring these structures given an object-oriented state repr
esentation, as well as a novel algorithm for Counterfactual Data Augmentation (C
oDA). CoDA uses local structures and an experience replay to generate counterfac
tual experiences that are causally valid in the global model. We find that CoDA
significantly improves the performance of RL agents in locally factored tasks, i
ncluding the batch-constrained and goal-conditioned settings. Code available at
https://github.com/spitis/mrl.
**********

Rethinking Learnable Tree Filter for Generic Feature Transform
Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Xiangyu Zhang, Hongbin Sun, Jian
 Sun, Nanning Zheng
The Learnable Tree Filter presents a remarkable approach to model structure-pres
erving relations for semantic segmentation. Nevertheless, the intrinsic geometri
c constraint forces it to focus on the regions with close spatial distance, hind
ering the effective long-range interactions. To relax the geometric constraint,
we give the analysis by reformulating it as a Markov Random Field and introduce
a learnable unary term. Besides, we propose a learnable spanning tree algorithm
to replace the original non-differentiable one, which further improves the flexi
bility and robustness. With the above improvements, our method can better captur
e long range dependencies and preserve structural details with linear complexity
, which is extended to several vision tasks for more generic feature transform.
Extensive experiments on object detection/instance segmentation demonstrate the
consistent improvements over the original version. For semantic segmentation, we
 achieve leading performance (82.1% mIoU) on the Cityscapes benchmark without be
lls-and whistles. Code is available at https://github.com/StevenGrove/LearnableT
reeFilterV2.
**********

Self-Supervised Relational Reasoning for Representation Learning
Massimiliano Patacchiola, Amos J. Storkey
In self-supervised learning, a system is tasked with achieving a surrogate objec
tive by defining alternative targets on a set of unlabeled data. The aim is to b
uild useful representations that can be used in downstream tasks, without costly
 manual annotation. In this work, we propose a novel self-supervised formulation
 of relational reasoning that allows a learner to bootstrap a signal from inform
ation implicit in unlabeled data. Training a relation head to discriminate how e
ntities relate to themselves (intra-reasoning) and other entities (inter-reasoni
ng), results in rich and descriptive representations in the underlying neural ne
twork backbone, which can be used in downstream tasks such as classification and
 image retrieval. We evaluate the proposed method following a rigorous experimen
tal procedure, using standard datasets, protocols, and backbones. Self-supervise

d relational reasoning outperforms the best competitor in all conditions by an a
verage 14% in accuracy, and the most recent state-of-the-art model by 3%. We lin
k the effectiveness of the method to the maximization of a Bernoulli log-likelih
ood, which can be considered as a proxy for maximizing the mutual information, r
esulting in a more efficient objective with respect to the commonly used contras
tive losses.
**********

Sufficient dimension reduction for classification using principal optimal transp
ort direction
Cheng Meng, Jun Yu, Jingyi Zhang, Ping Ma, Wenxuan Zhong
Sufficient dimension reduction is used pervasively as a supervised dimension red
uction approach.
Most existing sufficient dimension reduction methods are developed for data with
 a continuous response and may have an unsatisfactory performance for the catego
rical response, especially for the binary-response.
To address this issue, we propose a novel estimation method of sufficient dimens
ion reduction subspace (SDR subspace) using optimal transport.
The proposed method, named principal optimal transport direction (POTD), estimat
es the basis of the SDR subspace using the principal directions of the optimal t
ransport coupling between the data respecting different response categories.
The proposed method also reveals the relationship among three seemingly irreleva
nt topics, i.e., sufficient dimension reduction, support vector machine, and opt
imal transport.
We study the asymptotic properties of POTD and show that in the cases when the c
lass labels contain no error, POTD estimates the SDR subspace exclusively.
Empirical studies show POTD outperforms most of the state-of-the-art linear dime
nsion reduction methods.
**********

Fast Epigraphical Projection-based Incremental Algorithms for Wasserstein Distri
butionally Robust Support Vector Machine
Jiajin Li, Caihua Chen, Anthony Man-Cho So
Wasserstein \textbf{D}istributionally \textbf{R}obust \textbf{O}ptimization (DRO
) is concerned with finding decisions that perform well on data that are drawn f
rom the worst probability distribution within a Wasserstein ball centered at a c
ertain nominal distribution. In recent years, it has been shown that various DRO
 formulations of learning models admit tractable convex reformulations. However,
 most existing works propose to solve these convex reformulations by general-pur
pose solvers, which are not well-suited for tackling large-scale problems. In th
is paper, we focus on a family of Wasserstein distributionally robust support ve
ctor machine (DRSVM) problems and propose two novel epigraphical projection-base
d incremental algorithms to solve them. The updates in each iteration of these a
lgorithms can be computed in a highly efficient manner. Moreover, we show that t
he DRSVM problems considered in this paper satisfy a Hölderian growth condition
with explicitly determined growth exponents. Consequently, we are able to establ
ish the convergence rates of the proposed incremental algorithms. Our numerical
results indicate that the proposed methods are orders of magnitude faster than t
he state-of-the-art, and the performance gap grows considerably as the problem s
ize increases.
**********

Differentially Private Clustering: Tight Approximation Ratios
Badih Ghazi, Ravi Kumar, Pasin Manurangsi
We study the task of differentially private clustering.  For several basic clust
ering problems, including Euclidean DensestBall, 1-Cluster, k-means, and k-media
n, we give efficient differentially private algorithms that achieve essentially
the same approximation ratios as those that can be obtained by any non-private a
lgorithm, while incurring only small additive errors.  This improves upon existi
ng efficient algorithms that only achieve some large constant approximation fact
ors.
**********

On the Power of Louvain in the Stochastic Block Model

Vincent Cohen-Addad, Adrian Kosowski, Frederik Mallmann-Trenn, David Saulpic
A classic problem in machine learning and data analysis is to partition the vertices of a network in such a way that vertices in the same set are densely connected and vertices in different sets are loosely connected.
**********

Fairness with Overlapping Groups; a Probabilistic Perspective
Forest Yang, Mouhamadou Cisse, Sanmi Koyejo
In algorithmically fair prediction problems, a standard goal is to ensure the equality of fairness metrics across multiple overlapping groups simultaneously. We reconsider this standard fair classification problem using a probabilistic population analysis, which, in turn, reveals the Bayes-optimal classifier. Our approach unifies a variety of existing group-fair classification methods and enables extensions to a wide range of non-decomposable multiclass performance metrics and fairness measures. The Bayes-optimal classifier further inspires consistent procedures for algorithmically fair classification with overlapping groups. On a variety of real datasets, the proposed approach outperforms baselines in terms of its fairness-performance tradeoff.
**********

AttendLight: Universal Attention-Based Reinforcement Learning Model for Traffic Signal Control
Afshin Oroojlooy, Mohammadreza Nazari, Davood Hajinezhad, Jorge Silva
We propose AttendLight, an end-to-end Reinforcement Learning (RL) algorithm for the problem of traffic signal control. Previous approaches for this problem have the shortcoming that they require training for each new intersection with a different structure or traffic flow distribution. AttendLight solves this issue by training a single, universal model for intersections with any number of roads, lanes, phases (possible signals), and traffic flow. To this end, we propose a deep RL model which incorporates two attention models. The first attention model is introduced to handle different numbers of roads-lanes; and the second attention model is intended for enabling decision-making with any number of phases in an intersection. As a result, our proposed model works for any intersection configuration, as long as a similar configuration is represented in the training set. Experiments were conducted with both synthetic and real-world standard benchmark datasets. Our numerical experiment covers intersections with three or four approaching roads; one-directional/bi-directional roads with one, two, and three lanes; different number of phases; and different traffic flows. We consider two regimes: (i) single-environment training, single-deployment, and (ii) multi-environment training, multi-deployment. AttendLight outperforms both classical and other RL-based approaches on all cases in both regimes.
**********

Searching for Low-Bit Weights in Quantized Neural Networks
Zhaohui Yang, Yunhe Wang, Kai Han, Chunjing XU, Chao Xu, Dacheng Tao, Chang Xu
Quantized neural networks with low-bit weights and activations are attractive for developing AI accelerators. However, the quantization functions used in most conventional quantization methods are non-differentiable, which increases the optimization difficulty of quantized networks. Compared with full-precision parameters (\emph{i.e.}, 32-bit floating numbers), low-bit values are selected from a much smaller set. For example, there are only 16 possibilities in 4-bit space. Thus, we present to regard the discrete weights in an arbitrary quantized neural network as searchable variables, and utilize a differential method to search them accurately. In particular, each weight is represented as a probability distribution over the discrete value set. The probabilities are optimized during training and the values with the highest probability are selected to establish the desired quantized network. Experimental results on benchmarks demonstrate that the proposed method is able to produce quantized neural networks with higher performance over the state-of-the-arts on both image classification and super-resolution tasks.
**********

Adaptive Reduced Rank Regression
Qiong Wu, Felix MF Wong, Yanhua Li, Zhenming Liu, Varun Kanade

We study the low rank regression problem y = Mx + ε, where x and y are d1 and d2 dimensional vectors respectively. We consider the extreme high-dimensional setting where the number of observations n is less than d1 + d2. Existing algorithms are designed for settings where n is typically as large as rank(M)(d1+d2). This work provides an efficient algorithm which only involves two SVD, and establishes statistical guarantees on its performance. The algorithm decouples the problem by first estimating the precision matrix of the features, and then solving the matrix denoising problem. To complement the upper bound, we introduce new techniques for establishing lower bounds on the performance of any algorithm for this problem. Our preliminary experiments confirm that our algorithm often out-performs existing baseline, and is always at least competitive.

**********

From Predictions to Decisions: Using Lookahead Regularization

Nir Rosenfeld, Anna Hilgard, Sai Srivatsa Ravindranath, David C. Parkes

Machine learning is a powerful tool for predicting human-related outcomes, from creditworthiness to heart attack risks. But when deployed transparently, learned models also affect how users act in order to improve outcomes. The standard approach to learning predictive models is agnostic to induced user actions and provides no guarantees as to the effect of actions. We provide a framework for learning predictors that are accurate, while also considering interactions between the learned model and user decisions. For this, we introduce look-ahead regularization which, by anticipating user actions, encourages predictive models to also induce actions that improve outcomes. This regularization carefully tailors the uncertainty estimates that govern confidence in this improvement to the distribution of model-induced actions. We report the results of experiments on real and synthetic data that show the effectiveness of this approach.

**********

Sequential Bayesian Experimental Design with Variable Cost Structure

Sue Zheng, David Hayden, Jason Pacheco, John W. Fisher III

Mutual information (MI) is a commonly adopted utility function in Bayesian optimal experimental design (BOED). While theoretically appealing, MI evaluation poses a significant computational burden for most real world applications. As a result, many algorithms utilize MI bounds as proxies that lack regret-style guarantees. Here, we utilize two-sided bounds to provide such guarantees. Bounds are successively refined/tightened through additional computation until a desired guarantee is achieved. We consider the problem of adaptively allocating computational resources in BOED. Our approach achieves the same guarantee as existing methods, but with fewer evaluations of the costly MI reward. We adapt knapsack optimization of best arm identification problems, with important differences that impact overall algorithm design and performance. First, observations of MI rewards are biased. Second, evaluating experiments incurs shared costs amongst all experiments (posterior sampling) in addition to per experiment costs that may vary with increasing evaluation. We propose and demonstrate an algorithm that accounts for these variable costs in the refinement decision.

**********

Predictive inference is free with the jackknife+-after-bootstrap

Byol Kim, Chen Xu, Rina Barber

Ensemble learning is widely used in applications to make predictions in complex decision problems---for example, averaging models fitted to a sequence of samples bootstrapped from the available training data. While such methods offer more accurate, stable, and robust predictions and model estimates, much less is known about how to perform valid, assumption-lean inference on the output of these types of procedures. In this paper, we propose the jackknife+-after-bootstrap (J+aB), a procedure for constructing a predictive interval, which uses only the available bootstrapped samples and their corresponding fitted models, and is therefore "free" in terms of the cost of model fitting. The J+aB offers a predictive coverage guarantee that holds with no assumptions on the distribution of the data, the nature of the fitted model, or the way in which the ensemble of models are aggregated---at worst, the failure rate of the predictive interval is inflated by a factor of 2. Our numerical experiments verify the coverage and accuracy of th

e resulting predictive intervals on real data.
**********

Counterfactual Predictions under Runtime Confounding
Amanda Coston, Edward Kennedy, Alexandra Chouldechova
Algorithms are commonly used to predict outcomes under a particular decision or intervention, such as predicting likelihood of default if a loan is approved. Generally, to learn such counterfactual prediction models from observational data on historical decisions and corresponding outcomes, one must measure all factors that jointly affect the outcome and the decision taken.
Motivated by decision support applications, we study the counterfactual prediction task in the setting where all relevant factors are captured in the historical data, but it is infeasible, undesirable, or impermissible to use some such factors in the prediction model.
We refer to this setting as runtime confounding.
We propose a doubly-robust procedure for learning counterfactual prediction models in this setting.
Our theoretical analysis and experimental results suggest that our method often outperforms competing approaches.
We also present a validation procedure for evaluating the performance of counterfactual prediction methods.
**********

Learning Loss for Test-Time Augmentation
Ildoo Kim, Younghoon Kim, Sungwoong Kim
Data augmentation has been actively studied for robust neural networks. Most of the recent data augmentation methods focus on augmenting datasets during the training phase. At the testing phase, simple transformations are still widely used for test-time augmentation. This paper proposes a novel instance-level test- time augmentation that efficiently selects suitable transformations for a test input. Our proposed method involves an auxiliary module to predict the loss of each possible transformation given the input. Then, the transformations having lower predicted losses are applied to the input. The network obtains the results by averaging the prediction results of augmented inputs. Experimental results on several image classification benchmarks show that the proposed instance-aware test-time augmentation improves the model’s robustness against various corruptions.
**********

Balanced Meta-Softmax for Long-Tailed Visual Recognition
Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, hongsheng Li
Deep classifiers have achieved great success in visual recognition. However, real-world data is long-tailed by nature, leading to the mismatch between training and testing distributions. In this paper, we show that the Softmax function, though used in most classification tasks, gives a biased gradient estimation under the long-tailed setup. This paper presents Balanced Softmax, an elegant unbiased extension of Softmax, to accommodate the label distribution shift between training and testing. Theoretically, we derive the generalization bound for multiclass Softmax regression and show our loss minimizes the bound. In addition, we introduce Balanced Meta-Softmax, applying a complementary Meta Sampler to estimate the optimal class sample rate and further improve long-tailed learning. In our experiments, we demonstrate that Balanced Meta-Softmax outperforms state-of-the-art long-tailed classification solutions on both visual recognition and instance segmentation tasks.
**********

Efficient Exploration of Reward Functions in Inverse Reinforcement Learning via Bayesian Optimization
Sreejith Balakrishnan, Quoc Phong Nguyen, Bryan Kian Hsiang Low, Harold Soh
The problem of inverse reinforcement learning (IRL) is relevant to a variety of tasks including value alignment and robot learning from demonstration. Despite significant algorithmic contributions in recent years, IRL remains an ill-posed problem at its core; multiple reward functions coincide with the observed behavior and the actual reward function is not identifiable without prior knowledge or supplementary information. This paper presents an IRL framework called Bayesian

optimization-IRL (BO-IRL) which identifies multiple solutions that are consisten t with the expert demonstrations by efficiently exploring the reward function sp ace. BO-IRL achieves this by utilizing Bayesian Optimization along with our newl y proposed kernel that (a) projects the parameters of policy invariant reward fu nctions to a single point in a latent space and (b) ensures nearby points in the latent space correspond to reward functions yielding similar likelihoods. This projection allows the use of standard stationary kernels in the latent space to capture the correlations present across the reward function space. Empirical res ults on synthetic and real-world environments (model-free and model-based) show that BO-IRL discovers multiple reward functions while minimizing the number of e xpensive exact policy optimizations.
**********

MDP Homomorphic Networks: Group Symmetries in Reinforcement Learning
Elise van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, Max Welling
This paper introduces MDP homomorphic networks for deep reinforcement learning. MDP homomorphic networks are neural networks that are equivariant under symmetri es in the joint state-action space of an MDP. Current approaches to deep reinfor cement learning do not usually exploit knowledge about such structure. By buildi ng this prior knowledge into policy and value networks using an equivariance con straint, we can reduce the size of the solution space. We specifically focus on group-structured symmetries (invertible transformations). Additionally, we intro duce an easy method for constructing equivariant network layers numerically, so the system designer need not solve the constraints by hand, as is typically done . We construct MDP homomorphic MLPs and CNNs that are equivariant under either a group of reflections or rotations. We show that such networks converge faster t han unstructured baselines on CartPole, a grid world and Pong.
**********

How Can I Explain This to You? An Empirical Study of Deep Neural Network Explana tion Methods
Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, Mani Srivastava
Explaining the inner workings of deep neural network models have received consid erable attention in recent years. Researchers have attempted to provide human pa rseable explanations justifying why a model performed a specific classification. Although many of these toolkits are available for use, it is unclear which styl e of explanation is preferred by end-users, thereby demanding investigation. We performed a cross-analysis Amazon Mechanical Turk study comparing the popular st ate-of-the-art explanation methods to empirically determine which are better in explaining model decisions. The participants were asked to compare explanation m ethods across applications spanning image, text, audio, and sensory domains. Amo ng the surveyed methods, explanation-by-example was preferred in all domains exc ept text sentiment classification, where LIME's method of annotating input text was preferred. We highlight qualitative aspects of employing the studied explain ability methods and conclude with implications for researchers and engineers tha t seek to incorporate explanations into user-facing deployments.
**********

On the Error Resistance of Hinge-Loss Minimization
Kunal Talwar
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Munchausen Reinforcement Learning
Nino Vieillard, Olivier Pietquin, Matthieu Geist
Bootstrapping is a core mechanism in Reinforcement Learning (RL). Most algorithm s, based on temporal differences, replace the true value of a transiting state b y their current estimate of this value. Yet, another estimate could be leveraged to bootstrap RL: the current policy. Our core contribution stands in a very sim ple idea: adding the scaled log-policy to the immediate reward. We show that, by slightly modifying Deep Q-Network (DQN) in that way provides an agent that is c

ompetitive with the state-of-the-art Rainbow on Atari games, without making use of distributional RL, n-step returns or prioritized replay. To demonstrate the v ersatility of this idea, we also use it together with an Implicit Quantile Netwo rk (IQN). The resulting agent outperforms Rainbow on Atari, installing a new Sta te of the Art with very little modifications to the original algorithm. To add t o this empirical study, we provide strong theoretical insights on what happens u nder the hood -- implicit Kullback-Leibler regularization and increase of the ac tion-gap.
**********
Object Goal Navigation using Goal-Oriented Semantic Exploration
Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, Russ R. Salak hutdinov
This work studies the problem of object goal navigation which involves navigatin g to an instance of the given object category in unseen environments. End-to-end learning-based navigation methods struggle at this task as they are ineffective at exploration and long-term planning. We propose a modular system called, `Goa l-Oriented Semantic Exploration' which builds an episodic semantic map and uses it to explore the environment efficiently based on the goal object category. Emp irical results in visually realistic simulation environments show that the propo sed model outperforms a wide range of baselines including end-to-end learning-ba sed methods as well as modular map-based methods and led to the winning entry of the CVPR-2020 Habitat ObjectNav Challenge. Ablation analysis indicates that the proposed model learns semantic priors of the relative arrangement of objects in a scene, and uses them to explore efficiently. Domain-agnostic module design al lows us to transfer our model to a mobile robot platform and achieve similar per formance for object goal navigation in the real-world.
**********
Efficient semidefinite-programming-based inference for binary and multi-class MR Fs
Chirag Pabbaraju, Po-Wei Wang, J. Zico Kolter
Probabilistic inference in pairwise Markov Random Fields (MRFs), i.e. computing the partition function or computing a MAP estimate of the variables, is a founda tional problem in probabilistic graphical models. Semidefinite programming relax ations have long been a theoretically powerful tool for analyzing properties of probabilistic inference, but have not been practical owing to the high computati onal cost of typical solvers for solving the resulting SDPs. In this paper, we p ropose an efficient method for computing the partition function or MAP estimate in a pairwise MRF by instead exploiting a recently proposed coordinate-descent-b ased fast semidefinite solver. We also extend semidefinite relaxations from the typical binary MRF to the full multi-class setting, and develop a compact semide finite relaxation that can again be solved efficiently using the solver. We sho w that the method substantially outperforms (both in terms of solution quality a nd speed) the existing state of the art in approximate inference, on benchmark p roblems drawn from previous work. We also show that our approach can scale to la rge MRF domains such as fully-connected pairwise CRF models used in computer vis ion.
**********
Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language P rocessing
Zihang Dai, Guokun Lai, Yiming Yang, Quoc Le
With the success of language pretraining, it is highly desirable to develop more efficient architectures of good scalability that can exploit the abundant unlab eled data at a lower cost.
To improve the efficiency, we examine the much-overlooked redundancy in maintain ing a full-length token-level presentation, especially for tasks that only requi re a single-vector presentation of the sequence.
With this intuition, we propose Funnel-Transformer which gradually compresses th e sequence of hidden states to a shorter one and hence reduces the computation c ost.
More importantly, by re-investing the saved FLOPs from length reduction in const

ructing a deeper or wider model, we further improve the model capacity.
In addition, to perform token-level predictions as required by common pretrainin g objectives, Funnel-Transformer is able to recover a deep representation for ea ch token from the reduced hidden sequence via a decoder.
Empirically, with comparable or fewer FLOPs, Funnel-Transformer outperforms the standard Transformer on a wide variety of sequence-level prediction tasks, inclu ding text classification, language understanding, and reading comprehension.
**********

Semantic Visual Navigation by Watching YouTube Videos
Matthew Chang, Arjun Gupta, Saurabh Gupta
Semantic cues and statistical regularities in real-world environment layouts can improve efficiency for navigation in novel environments. This paper learns and leverages such semantic cues for navigating to objects of interest in novel envi ronments, by simply watching YouTube videos. This is challenging because YouTube videos don't come with labels for actions or goals, and may not even showcase o ptimal behavior. Our method tackles these challenges through the use of Q-learni ng on pseudo-labeled transition quadruples (image, action, next image, reward). We show that such off-policy Q-learning from passive data is able to learn meani ngful semantic cues for navigation. These cues, when used in a hierarchical navi gation policy, lead to improved efficiency at the ObjectGoal task in visually re alistic simulations. We observe a relative improvement of 15-83% over end-to-end RL, behavior cloning, and classical methods, while using minimal direct interac tion.
**********

Heavy-tailed Representations, Text Polarity Classification & Data Augmentation
Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emm anuel Vignon, Anne Sabourin
The dominant approaches to text representation in natural language rely on learn ing embeddings on massive corpora which have convenient properties such as compo sitionality and distance preservation. In this paper, we develop a novel method to learn a heavy-tailed embedding with desirable regularity properties regarding the distributional tails, which allows to analyze the points far away from the distribution bulk using the framework of multivariate extreme value theory. In p articular, a classifier dedicated to the tails of the proposed embedding is obta ined which exhibits a scale invariance property exploited in a novel text gener ation method for label preserving dataset augmentation. Experiments on synthetic and real text data show the relevance of the proposed framework and confirm tha t this method generates meaningful sentences with controllable attribute, e.g. p ositive or negative sentiments.
**********

SuperLoss: A Generic Loss for Robust Curriculum Learning
Thibault Castells, Philippe Weinzaepfel, Jerome Revaud
Curriculum learning is a technique to improve a model performance and generaliza tion based on the idea that easy samples should be presented before difficult on es during training. While it is generally complex to estimate a priori the diffi culty of a given sample, recent works have shown that curriculum learning can be formulated dynamically in a self-supervised manner. The key idea is to somehow estimate the importance (or weight) of each sample directly during training base d on the observation that easy and hard samples behave differently and can there fore be separated. However, these approaches are usually limited to a specific t ask (e.g., classification) and require extra data annotations, layers or paramet ers as well as a dedicated training procedure. We propose instead a simple and g eneric method that can be applied to a variety of losses and tasks without any c hange in the learning procedure. It consists in appending a novel loss function on top of any existing task loss, hence its name: the SuperLoss. Its main effect is to automatically downweight the contribution of samples with a large loss, i .e. hard samples, effectively mimicking the core principle of curriculum learnin g. As a side effect, we show that our loss prevents the memorization of noisy sa mples, making it possible to train from noisy data even with non-robust loss fun ctions. Experimental results on image classification, regression, object detecti

on and image retrieval demonstrate consistent gain, particularly in the presence of noise.
**********

CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models
Vijil Chenthamarakshan, Payel Das, Samuel Hoffman, Hendrik Strobelt, Inkit Padhi, Kar Wai Lim, Benjamin Hoover, Matteo Manica, Jannis Born, Teodoro Laino, Aleksandra Mojsilovic
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Memory Based Trajectory-conditioned Policies for Learning from Sparse Rewards
Yijie Guo, Jongwook Choi, Marcin Moczulski, Shengyu Feng, Samy Bengio, Mohammad Norouzi, Honglak Lee
Reinforcement learning with sparse rewards is challenging because an agent can rarely obtain non-zero rewards and hence, gradient-based optimization of parameterized policies can be incremental and slow. Recent work demonstrated that using a memory buffer of previous successful trajectories can result in more effective policies. However, existing methods may overly exploit past successful experiences, which can encourage the agent to adopt sub-optimal and myopic behaviors. In this work, instead of focusing on good experiences with limited diversity, we propose to learn a trajectory-conditioned policy to follow and expand diverse past trajectories from a memory buffer. Our method allows the agent to reach diverse regions in the state space and improve upon the past trajectories to reach new states. We empirically show that our approach significantly outperforms count-based exploration methods (parametric approach) and self-imitation learning (parametric approach with non-parametric memory) on various complex tasks with local optima. In particular, without using expert demonstrations or resetting to arbitrary states, we achieve the state-of-the-art scores under five billion number of frames, on challenging Atari games such as Montezuma’s Revenge and Pitfall.
**********

Liberty or Depth: Deep Bayesian Neural Nets Do Not Need Complex Weight Posterior Approximations
Sebastian Farquhar, Lewis Smith, Yarin Gal
We challenge the longstanding assumption that the mean-field approximation for variational inference in Bayesian neural networks is severely restrictive, and show this is not the case in deep networks. We prove several results indicating that deep mean-field variational weight posteriors can induce similar distributions in function-space to those induced by shallower networks with complex weight posteriors. We validate our theoretical contributions empirically, both through examination of the weight posterior using Hamiltonian Monte Carlo in small models and by comparing diagonal- to structured-covariance in large settings. Since complex variational posteriors are often expensive and cumbersome to implement, our results suggest that using mean-field variational inference in a deeper model is both a practical and theoretically justified alternative to structured approximations.
**********

Improving Sample Complexity Bounds for (Natural) Actor-Critic Algorithms
Tengyu Xu, Zhe Wang, Yingbin 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.
**********

Learning Differential Equations that are Easy to Solve
Jacob Kelly, Jesse Bettencourt, Matthew J. Johnson, David K. Duvenaud
Differential equations parameterized by neural networks become expensive to solve numerically as training progresses. We propose a remedy that encourages learne

d dynamics to be easier to solve. Specifically, we introduce a differentiable su rrogate for the time cost of standard numerical solvers, using higher-order deri vatives of solution trajectories. These derivatives are efficient to compute wit h Taylor-mode automatic differentiation. Optimizing this additional objective tr ades model performance against the time cost of solving the learned dynamics. We demonstrate our approach by training substantially faster, while nearly as accu rate, models in supervised classification, density estimation, and time-series m odelling tasks.
**********

Stability of Stochastic Gradient Descent on Nonsmooth Convex Losses
Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, Kunal Talwar
Uniform stability is a notion of algorithmic stability that bounds the worst cas e change in the model output by the algorithm when a single data point in the da taset is replaced. An influential work of Hardt et al. [2016] provides strong up per bounds on the uniform stability of the stochastic gradient descent (SGD) alg orithm on sufficiently smooth convex losses. These results led to important prog ress in understanding of the generalization properties of SGD and several applic ations to differentially private convex optimization for smooth losses.
**********

Influence-Augmented Online Planning for Complex Environments
Jinke He, Miguel Suau de Castro, Frans Oliehoek
How can we plan efficiently in real time to control an agent in a complex enviro nment that may involve many other agents? While existing sample-based planners h ave enjoyed empirical success in large POMDPs, their performance heavily relies on a fast simulator. However, real-world scenarios are complex in nature and the ir simulators are often computationally demanding, which severely limits the per formance of online planners. In this work, we propose influence-augmented online planning, a principled method to transform a factored simulator of the entire e nvironment into a local simulator that samples only the state variables that are most relevant to the observation and reward of the planning agent and captures the incoming influence from the rest of the environment using machine learning m ethods. Our main experimental results show that planning on this less accurate b ut much faster local simulator with POMCP leads to higher real-time planning per formance than planning on the simulator that models the entire environment.
**********

PAC-Bayes Learning Bounds for Sample-Dependent Priors
Pranjal Awasthi, Satyen Kale, Stefani Karp, Mehryar Mohri
We present a series of new PAC-Bayes learning guarantees for randomized algorith ms with sample-dependent priors. Our most general bounds make no assumption on t he priors and are given in terms of certain covering numbers under the infinite-Renyi divergence and the L1 distance. We show how to use these general bounds to derive leaning bounds in the setting where the sample-dependent priors obey an infinite-Renyi divergence or L1-distance sensitivity condition. We also provide a flexible framework for computing PAC-Bayes bounds, under certain stability ass umptions on the sample-dependent priors, and show how to use this framework to g ive more refined bounds when the priors satisfy an infinite-Renyi divergence sen sitivity condition.
**********

Reward-rational (implicit) choice: A unifying formalism for reward learning
Hong Jun Jeon, Smitha Milli, Anca Dragan
It is often difficult to hand-specify what the correct reward function is for a task, so researchers have instead aimed to learn reward functions from human beh avior or feedback. The types of behavior interpreted as evidence of the reward f unction have expanded greatly in recent years. We've gone from demonstrations, t o comparisons, to reading into the information leaked when the human is pushing the robot away or turning it off. And surely, there is more to come. How will a robot make sense of all these diverse types of behavior? Our key observation is that different types of behavior can be interpreted in a single unifying formali sm - as a reward-rational choice that the human is making, often implicitly. We use this formalism to survey prior work through a unifying lens, and discuss its

potential use as a recipe for interpreting new sources of information that are yet to be uncovered.
**********
Probabilistic Time Series Forecasting with Shape and Temporal Diversity
Vincent LE GUEN, Nicolas THOME
Probabilistic forecasting consists in predicting a distribution of possible futu re outcomes. In this paper, we address this problem for non-stationary time seri es, which is very challenging yet crucially important. We introduce the STRIPE m odel for representing structured diversity based on shape and time features, ens uring both probable predictions while being sharp and accurate. STRIPE is agnost ic to the forecasting model, and we equip it with a diversification mechanism re lying on determinantal point processes (DPP). We introduce two DPP kernels for m odelling diverse trajectories in terms of shape and time, which are both differe ntiable and proved to be positive semi-definite. To have an explicit control on the diversity structure, we also design an iterative sampling mechanism to disen tangle shape and time representations in the latent space. Experiments carried o ut on synthetic datasets show that STRIPE significantly outperforms baseline met hods for representing diversity, while maintaining accuracy of the forecasting m odel. We also highlight the relevance of the iterative sampling scheme and the i mportance to use different criteria for measuring quality and diversity. Finally , experiments on real datasets illustrate that STRIPE is able to outperform stat e-of-the-art probabilistic forecasting approaches in the best sample prediction.
**********
Low Distortion Block-Resampling with Spatially Stochastic Networks
Sarah Hong, Martin Arjovsky, Darryl Barnhart, Ian Thompson
We formalize and attack the problem of generating new images from old ones that are as diverse as possible, only allowing them to change without restrictions in certain parts of the image while remaining globally consistent.
This encompasses the typical situation found in generative modelling, where we a re happy with parts of the generated data, but would like to resample others (`` I like this generated castle overall, but this tower looks unrealistic, I would like a new one'').
In order to attack this problem we build from the best conditional and unconditi onal generative models to introduce a new network architecture, training procedu re,  and a new algorithm for resampling parts of the image as desired.
**********
Continual Deep Learning by Functional Regularisation of Memorable Past
Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner , Mohammad Emtiyaz E. Khan
Continually learning new skills is important for intelligent systems, yet standa rd deep learning methods suffer from catastrophic forgetting of the past. Recent works address this with weight regularisation. Functional regularisation, altho ugh computationally expensive, is expected to perform better, but rarely does so in practice. In this paper, we fix this issue by using a new functional-regular isation approach that utilises a few memorable past examples crucial to avoid fo rgetting. By using a Gaussian Process formulation of deep networks, our approach enables training in weight-space while identifying both the memorable past and a functional prior. Our method achieves state-of-the-art performance on standard benchmarks and opens a new direction for life-long learning where regularisatio n and memory-based methods are naturally combined.
**********
Distance Encoding: Design Provably More Powerful Neural Networks for Graph Repre sentation Learning
Pan Li, Yanbang Wang, Hongwei Wang, Jure Leskovec
Learning representations of sets of nodes in a graph is crucial for applications ranging from node-role discovery to link prediction and molecule classification . Graph Neural Networks (GNNs) have achieved great success in graph representati on learning. However, expressive power of GNNs is limited by the 1-Weisfeiler-Le hman (WL) test and thus GNNs generate identical representations for graph substr uctures that may in fact be very different. More powerful GNNs, proposed recentl

y by mimicking higher-order-WL tests, only focus on representing entire graphs a
nd they are computationally inefficient as they cannot utilize sparsity of the u
nderlying graph. Here we propose and mathematically analyze a general class of s
tructure-related features, termed Distance Encoding (DE). DE assists GNNs in rep
resenting any set of nodes, while providing strictly more expressive power than
the 1-WL test. DE captures the distance between the node set whose representatio
n is to be learned and each node in the graph. To capture the distance DE can ap
ply various graph-distance measures such as shortest path distance or generalize
d PageRank scores. We propose two ways for GNNs to use DEs (1) as extra node fea
tures, and (2) as controllers of message aggregation in GNNs. Both approaches ca
n utilize the sparse structure of the underlying graph, which leads to computati
onal efficiency and scalability. We also prove that DE can distinguish node sets
 embedded in almost all regular graphs where traditional GNNs always fail. We ev
aluate DE on three tasks over six real networks: structural role prediction, lin
k prediction, and triangle prediction. Results show that our models outperform G
NNs without DE by up-to 15\% in accuracy and AUROC. Furthermore, our models also
 significantly outperform other state-of-the-art methods especially designed for
 the above tasks.
**********

Fast Fourier Convolution
Lu Chi, Borui Jiang, Yadong Mu
Vanilla convolutions in modern deep networks are known to operate locally and at
 fixed scale (e.g., the widely-adopted 3*3 kernels in image-oriented tasks). Thi
s causes low efficacy in connecting two distant locations in the network. In thi
s work, we propose a novel convolutional operator dubbed as fast Fourier convolu
tion (FFC), which has the main hallmarks of non-local receptive fields and cross
-scale fusion within the convolutional unit. According to spectral convolution t
heorem in Fourier theory, point-wise update in the spectral domain globally affe
cts all input features involved in Fourier transform, which sheds light on neura
l architectural design with non-local receptive field. Our proposed FFC is inspi
red to capsulate three different kinds of computations in a single operation uni
t: a local branch that conducts ordinary small-kernel convolution, a semi-global
 branch that processes spectrally stacked image patches, and a global branch tha
t manipulates image-level spectrum. All branches complementarily address differe
nt scales. A multi-branch aggregation step is included in FFC for cross-scale fu
sion. FFC is a generic operator that can directly replace vanilla convolutions i
n a large body of existing networks, without any adjustments and with comparable
 complexity metrics (e.g., FLOPs). We experimentally evaluate FFC in three major
 vision benchmarks (ImageNet for image recognition, Kinetics for video action re
cognition, MSCOCO for human keypoint detection). It consistently elevates accura
cies in all above tasks by significant margins.
**********

Unsupervised Learning of Dense Visual Representations
Pedro O. O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, Aaron C. C
ourville
Contrastive self-supervised learning has emerged as a promising approach to unsu
pervised visual representation learning. In general, these methods learn global
(image-level)  representations that are invariant to different views (i.e., comp
ositions of data augmentation) of the same image. However, many visual understan
ding tasks require dense (pixel-level) representations. In this paper, we propos
e View-Agnostic Dense Representation (VADeR) for unsupervised learning of dense
representations. VADeR learns pixelwise representations by forcing local feature
s to remain constant over different viewing conditions. Specifically, this is ac
hieved through pixel-level contrastive learning: matching features (that is, fea
tures that describes the same location of the scene on different views) should b
e close in an embedding space, while non-matching features should be apart. VADe
R provides a natural representation for dense prediction tasks and transfers wel
l to downstream tasks. Our method outperforms ImageNet supervised pretraining (a
nd strong unsupervised baselines) in multiple dense prediction tasks.
**********

Higher-Order Certification For Randomized Smoothing
Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, Luca Daniel
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Structured Distributions From Untrusted Batches: Faster and Simpler
Sitan Chen, Jerry Li, Ankur Moitra
We revisit the problem of learning from untrusted batches introduced by Qiao and Valiant [QV17]. Recently, Jain and Orlitsky [JO19] gave a simple semidefinite programming approach based on the cut-norm that achieves essentially information-theoretically optimal error in polynomial time. Concurrently, Chen et al. [CLM19] considered a variant of the problem where $\mu$ is assumed to be structured, e.g. log-concave, monotone hazard rate, t-modal, etc. In this case, it is possible to achieve the same error with sample complexity sublinear in n, and they exhibited a quasi-polynomial time algorithm for doing so using Haar wavelets.
**********

Hierarchical Quantized Autoencoders
Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, John Hughes
Despite progress in training neural networks for lossy image compression, current approaches fail to maintain both perceptual quality and abstract features at very low bitrates. Encouraged by recent success in learning discrete representations with Vector Quantized Variational Autoencoders (VQ-VAEs), we motivate the use of a hierarchy of VQ-VAEs to attain high factors of compression. We show that the combination of stochastic quantization and hierarchical latent structure aids likelihood-based image compression. This leads us to introduce a novel objective for training hierarchical VQ-VAEs. Our resulting scheme produces a Markovian series of latent variables that reconstruct images of high-perceptual quality which retain semantically meaningful features. We provide qualitative and quantitative evaluations on the CelebA and MNIST datasets.
**********

Diversity can be Transferred: Output Diversification for White- and Black-box Attacks
Yusuke Tashiro, Yang Song, Stefano Ermon
Adversarial attacks often involve random perturbations of the inputs drawn from uniform or Gaussian distributions, e.g. to initialize optimization-based white-box attacks or generate update directions in black-box attacks. These simple perturbations, however, could be sub-optimal as they are agnostic to the model being attacked. To improve the efficiency of these attacks, we propose Output Diversified Sampling (ODS), a novel sampling strategy that attempts to maximize diversity in the target model's outputs among the generated samples. While ODS is a gradient-based strategy, the diversity offered by ODS is transferable and can be helpful for both white-box and black-box attacks via surrogate models. Empirically, we demonstrate that ODS significantly improves the performance of existing white-box and black-box attacks. In particular, ODS reduces the number of queries needed for state-of-the-art black-box attacks on ImageNet by a factor of two.
**********

POLY-HOOT: Monte-Carlo Planning in Continuous Space MDPs with Non-Asymptotic Analysis
Weichao Mao, Kaiqing Zhang, Qiaomin Xie, Tamer Basar
Monte-Carlo planning, as exemplified by Monte-Carlo Tree Search (MCTS), has demonstrated remarkable performance in applications with finite spaces. In this paper, we consider Monte-Carlo planning in an environment with continuous state-action spaces, a much less understood problem with important applications in control and robotics. We introduce POLY-HOOT, an algorithm that augments MCTS with a continuous armed bandit strategy named Hierarchical Optimistic Optimization (HOO) (Bubeck et al., 2011). Specifically, we enhance HOO by using an appropriate polynomial, rather than logarithmic, bonus term in the upper confidence bounds. Such a polynomial bonus is motivated by its empirical successes in AlphaGo Zero (Sil

ver et al., 2017b), as well as its significant role in achieving theoretical gua rantees of finite space MCTS (Shah et al., 2019). We investigate, for the first time, the regret of the enhanced HOO algorithm in non-stationary bandit problems . Using this result as a building block, we establish non-asymptotic convergence guarantees for POLY-HOOT: the value estimate converges to an arbitrarily small neighborhood of the optimal value function at a polynomial rate. We further prov ide experimental results that corroborate our theoretical findings.
**********

AvE: Assistance via Empowerment

Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, Anca Dragan

One difficulty in using artificial agents for human-assistive applications lies in the challenge of accurately assisting with a person's goal(s). Existing meth ods tend to rely on inferring the human's goal, which is challenging when there are many potential goals or when the set of candidate goals is difficult to iden tify. We propose a new paradigm for assistance by instead increasing the human's ability to control their environment, and formalize this approach by augmenting reinforcement learning with human empowerment. This task-agnostic objective inc reases the person's autonomy and ability to achieve any eventual state. We test our approach against assistance based on goal inference, highlighting scenarios where our method overcomes failure modes stemming from goal ambiguity or misspec ification. As existing methods for estimating empowerment in continuous domains are computationally hard, precluding its use in real time learned assistance, we also propose an efficient empowerment-inspired proxy metric. Using this, we are able to successfully demonstrate our method in a shared autonomy user study for a challenging simulated teleoperation task with human-in-the-loop training.
**********

Variational Policy Gradient Method for Reinforcement Learning with General Utili ties

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, Mengdi Wang

In recent years, reinforcement learning systems with general goals beyond a cumu lative sum of rewards have gained traction, such as in constrained problems, exp loration, and acting upon prior experiences. In this paper, we consider policy o ptimization in Markov Decision Problems, where the objective is a general utilit y function of the state-action occupancy measure, which subsumes several of the aforementioned examples as special cases. Such generality invalidates the Bellma n equation. As this means that dynamic programming no longer works, we focus on direct policy search. Analogously to the Policy Gradient Theorem \cite{sutton200 0policy} available for RL with cumulative rewards, we derive a new Variational P olicy Gradient Theorem for RL with general utilities, which establishes that the gradient may be obtained as the solution of a stochastic saddle point problem i nvolving the Fenchel dual of the utility function. We develop a variational Mont e Carlo gradient estimation algorithm to compute the policy gradient based on sa mple paths. Further, we prove that the variational policy gradient scheme conver ges globally to the optimal policy for the general objective, and we also establ ish its rate of convergence that matches or improves the convergence rate availa ble in the case of RL with cumulative rewards.
**********

Reverse-engineering recurrent neural network solutions to a hierarchical inferen ce task for mice

Rylan Schaeffer, Mikail Khona, Leenoy Meshulam, Brain Laboratory International, Ila Fiete

We study how recurrent neural networks (RNNs) solve a hierarchical inference tas k involving two latent variables and disparate timescales separated by 1-2 order s of magnitude. The task is of interest to the International Brain Laboratory, a global collaboration of experimental and theoretical neuroscientists studying h ow the mammalian brain generates behavior. We make four discoveries. First, RNNs learn behavior that is quantitatively similar to ideal Bayesian baselines. Seco nd, RNNs perform inference by learning a two-dimensional subspace defining belie fs about the latent variables. Third, the geometry of RNN dynamics reflects an i nduced coupling between the two separate inference processes necessary to solve

the task. Fourth, we perform model compression through a novel form of knowledge distillation on hidden representations -- Representations and Dynamics Distillation (RADD)-- to reduce the RNN dynamics to a low-dimensional, highly interpretable model. This technique promises a useful tool for interpretability of high dimensional nonlinear dynamical systems. Altogether, this work yields predictions to guide exploration and analysis of mouse neural data and circuity.
**********

Temporal Positive-unlabeled Learning for Biomedical Hypothesis Generation via Risk Estimation
Uchenna Akujuobi, Jun Chen, Mohamed Elhoseiny, Michael Spranger, Xiangliang Zhang
Understanding the relationships between biomedical terms like viruses, drugs, and symptoms is essential in the fight against diseases. Many attempts have been made to introduce the use of machine learning to the scientific process of hypothesis generation (HG), which refers to the discovery of meaningful implicit connections between biomedical terms. However, most existing methods fail to truly capture the temporal dynamics of scientific term relations and also assume unobserved connections to be irrelevant (i.e., in a positive-negative (PN) learning setting). To break these limits, we formulate this HG problem as future connectivity prediction task on a dynamic attributed graph via positive-unlabeled (PU) learning. Then, the key is to capture the temporal evolution of node pair (term pair) relations from just the positive and unlabeled data. We propose a variational inference model to estimate the positive prior, and incorporate it in the learning of node pair embeddings, which are then used for link prediction. Experiment results on real-world biomedical term relationship datasets and case study analyses on a COVID-19 dataset validate the effectiveness of the proposed model.
**********

Efficient Low Rank Gaussian Variational Inference for Neural Networks
Marcin Tomczak, Siddharth Swaroop, Richard Turner
Bayesian neural networks are enjoying a renaissance driven in part by recent advances in variational inference (VI).
The most common form of VI employs a fully factorized or mean-field distribution, but this is known to suffer from several pathologies, especially as we expect posterior distributions with highly correlated parameters.
Current algorithms that capture these correlations with a Gaussian approximating family are difficult to scale to large models due to computational costs and high variance of gradient updates.
By using a new form of the reparametrization trick, we derive a computationally efficient algorithm for performing VI with a Gaussian family with a low-rank plus diagonal covariance structure.
We scale to deep feed-forward and convolutional architectures.
We find that adding low-rank terms to parametrized diagonal covariance does not improve predictive performance except on small networks, but low-rank terms added to a constant diagonal covariance improves performance on small and large-scale network architectures.
**********

Privacy Amplification via Random Check-Ins
Borja Balle, Peter Kairouz, Brendan McMahan, Om Thakkar, Abhradeep Guha Thakurta
Differentially Private Stochastic Gradient Descent (DP-SGD) forms a fundamental building block in many applications for learning over sensitive data. Two standard approaches, privacy amplification by subsampling, and privacy amplification by shuffling, permit adding lower noise in DP-SGD than via na\"{\i}ve schemes. A key assumption in both these approaches is that the elements in the data set can be uniformly sampled, or be uniformly permuted --- constraints that may become prohibitive when the data is processed in a decentralized or distributed fashio

n. In this paper, we focus on conducting iterative methods like DP-SGD in the se
tting of federated learning (FL) wherein the data is distributed among many devi
ces (clients). Our main contribution is the \emph{random check-in} distributed p
rotocol, which crucially relies only on randomized participation decisions made
locally and independently by each client. It has privacy/accuracy trade-offs sim
ilar to privacy amplification by subsampling/shuffling. However, our method does
not require server-initiated communication, or even knowledge of the population
size. To our knowledge, this is the first privacy amplification tailored for a
distributed learning framework, and it may have broader applicability beyond FL.
Along the way, we improve the privacy guarantees of amplification by shuffling
and show that, in practical regimes, this improvement allows for similar privacy
and utility using data from an order of magnitude fewer users.
**********

Probabilistic Circuits for Variational Inference in Discrete Graphical Models
Andy Shih, Stefano Ermon
Inference in discrete graphical models with variational methods is difficult bec
ause of the inability to re-parameterize gradients of the Evidence Lower Bound (
ELBO). Many sampling-based methods have been proposed for estimating these gradi
ents, but they suffer from high bias or variance. In this paper, we propose a ne
w approach that leverages the tractability of probabilistic circuit models, such
as Sum Product Networks (SPN), to compute ELBO gradients exactly (without sampl
ing) for a certain class of densities. In particular, we show that selective-SPN
s are suitable as an expressive variational distribution, and prove that when th
e log-density of the target model is a polynomial the corresponding ELBO can be
computed analytically. To scale to graphical models with thousands of variables,
we develop an efficient and effective construction of selective-SPNs with size
(O(kn)), where (n) is the number of variables and (k) is an adjustable hyperpara
meter. We demonstrate our approach on three types of graphical models -- Ising m
odels, Latent Dirichlet Allocation, and factor graphs from the UAI Inference Com
petition. Selective-SPNs give a better lower bound than mean-field and structure
d mean-field, and is competitive with approximations that do not provide a lower
bound, such as Loopy Belief Propagation and Tree-Reweighted Belief Propagation.
Our results show that probabilistic circuits are promising tools for variationa
l inference in discrete graphical models as they combine tractability and expres
sivity.
**********

Your Classifier can Secretly Suffice Multi-Source Domain Adaptation
Naveen Venkat, Jogendra Nath Kundu, Durgesh Singh, Ambareesh Revanur, Venkatesh
Babu R
Multi-Source Domain Adaptation (MSDA) deals with the transfer of task knowledge
from multiple labeled source domains to an unlabeled target domain, under a doma
in-shift. Existing methods aim to minimize this domain-shift using auxiliary dis
tribution alignment objectives. In this work, we present a different perspective
to MSDA wherein deep models are observed to implicitly align the domains under
label supervision. Thus, we aim to utilize implicit alignment without additional
training objectives to perform adaptation. To this end, we use pseudo-labeled t
arget samples and enforce a classifier agreement on the pseudo-labels, a process
called Self-supervised Implicit Alignment (SImpAl). We find that SImpAl readily
works even under category-shift among the source domains. Further, we propose c
lassifier agreement as a cue to determine the training convergence, resulting in
a simple training algorithm. We provide a thorough evaluation of our approach o
n five benchmarks, along with detailed insights into each component of our appro
ach.
**********

Labelling unlabelled videos from scratch with multi-modal self-supervision
Yuki Asano, Mandela Patrick, Christian Rupprecht, Andrea Vedaldi
A large part of the current success of deep learning lies in the effectiveness o
f data -- more precisely: of labeled data. Yet, labelling a dataset with human a
nnotation continues to carry high costs, especially for videos. While in the ima
ge domain, recent methods have allowed to generate meaningful (pseudo-) labels f

or unlabelled datasets without supervision, this development is missing for the video domain where learning feature representations is the current focus. In this work, we a) show that unsupervised labelling of a video dataset does not come for free from strong feature encoders and b) propose a novel clustering method that allows pseudo-labelling of a video dataset without any human annotations, by leveraging the natural correspondence between audio and visual modalities. An extensive analysis shows that the resulting clusters have high semantic overlap to ground truth human labels. We further introduce the first benchmarking results on unsupervised labelling of common video datasets.
**********

A Non-Asymptotic Analysis for Stein Variational Gradient Descent
Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, Arthur Gretton
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Robust Meta-learning for Mixed Linear Regression with Small Batches
Weihao Kong, Raghav Somani, Sham Kakade, 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.
**********

Bayesian Deep Learning and a Probabilistic Perspective of Generalization
Andrew G. Wilson, Pavel Izmailov
The key distinguishing property of a Bayesian approach is marginalization, rather than using a single setting of weights. Bayesian marginalization can particularly improve the accuracy and calibration of modern deep neural networks, which are typically underspecified by the data, and can represent many compelling but different solutions. We show that deep ensembles provide an effective mechanism for approximate Bayesian marginalization, and propose a related approach that further improves the predictive distribution by marginalizing within basins of attraction, without significant overhead. We also investigate the prior over functions implied by a vague distribution over neural network weights, explaining the generalization properties of such models from a probabilistic perspective. From this perspective, we explain results that have been presented as mysterious and distinct to neural network generalization, such as the ability to fit images with random labels, and show that these results can be reproduced with Gaussian processes. We also show that Bayesian model averaging alleviates double descent, resulting in monotonic performance improvements with increased flexibility.
**********

Unsupervised Learning of Object Landmarks via Self-Training Correspondence
Dimitrios Mallis, Enrique Sanchez, Matthew Bell, Georgios Tzimiropoulos
This paper addresses the problem of unsupervised discovery of object landmarks. We take a different path compared to that of existing works, based on 2 novel perspectives: (1) Self-training: starting from generic keypoints, we propose a self-training approach where the goal is to learn a detector that improves itself becoming more and more tuned to object landmarks. (2) Correspondence: we identify correspondence as a key objective for unsupervised landmark discovery and propose an optimization scheme which alternates between recovering object landmark correspondence across different images via clustering and learning an object landmark descriptor without labels. Compared to previous works, our approach can learn landmarks that are more flexible in terms of capturing large changes in viewpoint. We show the favourable properties of our method on a variety of difficult datasets including LS3D, BBCPose and Human3.6M. Code is available at https://github.com/malldimi1/UnsupervisedLandmarks
**********

Randomized tests for high-dimensional regression: A more efficient and powerful solution

Yue Li, Ilmun Kim, Yuting Wei
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Representations from Audio-Visual Spatial Alignment
Pedro Morgado, Yi Li, Nuno Nvasconcelos
We introduce a novel self-supervised pretext task for learning representations from audio-visual content. Prior work on audio-visual representation learning leverages correspondences at the video level. Approaches based on audio-visual correspondence (AVC) predict whether audio and video clips originate from the same or different video instances. Audio-visual temporal synchronization (AVTS) further discriminates negative pairs originated from the same video instance but at different moments in time. While these approaches learn high-quality representations for downstream tasks such as action recognition, they completely disregard the spatial cues of audio and visual signals naturally occurring in the real world. To learn from these spatial cues, we tasked a network to perform contrastive audio-visual spatial alignment of 360\degree video and spatial audio. The ability to perform spatial alignment is enhanced by reasoning over the full spatial content of the 360\degree video using a transformer architecture to combine representations from multiple viewpoints. The advantages of the proposed pretext task are demonstrated on a variety of audio and visual downstream tasks, including audio-visual correspondence, spatial alignment, action recognition and video semantic segmentation. Dataset and code are available at https://github.com/pedro-morgado/AVSpatialAlignment.
**********

Generative View Synthesis: From Single-view Semantics to Novel-view Images
Tewodros Amberbir Habtegebrial, Varun Jampani, Orazio Gallo, Didier Stricker
Content creation, central to applications such as virtual reality, can be tedious and time-consuming. Recent image synthesis methods simplify this task by offering tools to generate new views from as little as a single input image, or by converting a semantic map into a photorealistic image. We propose to push the envelope further, and introduce Generative View Synthesis (GVS) that can synthesize multiple photorealistic views of a scene given a single semantic map.
We show that the sequential application of existing techniques, e.g., semantics-to-image translation followed by monocular view synthesis, fail at capturing the scene's structure. In contrast, we solve the semantics-to-image translation in concert with the estimation of the 3D layout of the scene, thus producing geometrically consistent novel views that preserve semantic structures. We first lift the input 2D semantic map onto a 3D layered representation of the scene in feature space, thereby preserving the semantic labels of 3D geometric structures. We then project the layered features onto the target views to generate the final novel-view images. We verify the strengths of our method and compare it with several advanced baselines on three different datasets. Our approach also allows for style manipulation and image editing operations, such as the addition or removal of objects, with simple manipulations of the input style images and semantic maps respectively. For code and additional results, visit the project page at https://gvsnet.github.io
**********

Towards More Practical Adversarial Attacks on Graph Neural Networks
Jiaqi Ma, Shuangrui Ding, Qiaozhu Mei
We study the black-box attacks on graph neural networks (GNNs) under a novel and realistic constraint: attackers have access to only a subset of nodes in the network, and they can only attack a small number of them. A node selection step is essential under this setup. We demonstrate that the structural inductive biases of GNN models can be an effective source for this type of attacks. Specifically, by exploiting the connection between the backward propagation of GNNs and random walks, we show that the common gradient-based white-box attacks can be generalized to the black-box setting via the connection between the gradient and an im

portance score similar to PageRank. In practice, we find attacks based on this i
mportance score indeed increase the classification loss by a large margin, but t
hey fail to significantly increase the mis-classification rate. Our theoretical
and empirical analyses suggest that there is a discrepancy between the loss and
mis-classification rate, as the latter presents a diminishing-return pattern whe
n the number of attacked nodes increases. Therefore, we propose a greedy procedu
re to correct the importance score that takes into account of the diminishing-re
turn pattern. Experimental results show that the proposed procedure can signific
antly increase the mis-classification rate of common GNNs on real-world data wit
hout access to model parameters nor predictions.
**********

Multi-Task Reinforcement Learning with Soft Modularization
Ruihan Yang, Huazhe Xu, YI WU, Xiaolong Wang
Multi-task learning is a very challenging problem in reinforcement learning. Whi
le training multiple tasks jointly allow the policies to share parameters across
 different tasks, the optimization problem becomes non-trivial:  It remains uncl
ear what parameters in the network should be reused across tasks, and how the gr
adients from different tasks may interfere with each other. Thus, instead of nai
vely sharing parameters across tasks, we introduce an explicit modularization te
chnique on policy representation to alleviate this optimization issue. Given a b
ase policy network, we design a routing network which estimates different routin
g strategies to reconfigure the base network for each task. Instead of directly
selecting routes for each task, our task-specific policy uses a method called so
ft modularization to softly combine all the possible routes, which makes it suit
able for sequential tasks. We experiment with various robotics manipulation task
s in simulation and show our method improves both sample efficiency and performa
nce over strong baselines by a large margin.
**********

Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predict
ions of Complex Models
Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, Tom Claassen
Shapley values underlie one of the most popular model-agnostic methods within ex
plainable artificial intelligence. These values are designed to attribute the di
fference between a model's prediction and an average baseline to the different f
eatures used as input to the model. Being based on solid game-theoretic principl
es, Shapley values uniquely satisfy several desirable properties, which is why t
hey are increasingly used to explain the predictions of possibly complex and hig
hly non-linear machine learning models. Shapley values are well calibrated to a
user’s intuition when features are independent, but may lead to undesirable, cou
nterintuitive explanations when the independence assumption is violated.
**********

On the training dynamics of deep networks with $L_2$ regularization
Aitor Lewkowycz, Guy Gur-Ari
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Improved Algorithms for Convex-Concave Minimax Optimization
Yuanhao Wang, Jian 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.
**********

Deep Variational Instance Segmentation
Jialin Yuan, Chao Chen, Fuxin Li
Instance segmentation, which seeks to obtain both class and instance labels for
each pixel in the input image, is a challenging task in computer vision. State-o
f- the-art algorithms often employ a search-based strategy, which first divides

the output image with a regular grid and generate proposals at each grid cell, t
hen the proposals are classified and boundaries refined. In this paper, we propo
se a novel algorithm that directly utilizes a fully convolutional network (FCN)
to predict instance labels. Specifically, we propose a variational relaxation of
 instance segmentation as minimizing an optimization functional for a piecewise-
constant segmentation problem, which can be used to train an FCN end-to-end. It
extends the classical Mumford-Shah variational segmentation algorithm to be able
 to handle the permutation-invariant ground truth in instance segmentation. Expe
riments on PASCAL VOC 2012 and the MSCOCO 2017 dataset show that the proposed ap
proach efficiently tackles the instance segmentation task.
**********

Learning Implicit Functions for Topology-Varying Dense 3D Shape Correspondence
Feng Liu, Xiaoming Liu
The goal of this paper is to learn dense 3D shape correspondence for topology-va
rying objects in an unsupervised manner. Conventional implicit functions estimat
e the occupancy of a 3D point given a shape latent code. Instead, our novel impl
icit function produces a part embedding vector for each 3D point, which is assum
ed to be similar to its densely corresponded point in another 3D shape of the sa
me object category. Furthermore, we implement dense correspondence through an in
verse function mapping from the part embedding to a  corresponded 3D point. Both
 functions are jointly learned with several effective loss functions to realize
our assumption, together with the encoder generating the shape latent code.  Dur
ing inference, if a user selects an arbitrary point on the source shape, our alg
orithm can automatically generate a confidence score indicating whether there is
 a correspondence on the target shape, as well as the corresponding semantic poi
nt if there is. Such a mechanism inherently benefits man-made objects with diffe
rent part constitutions. The effectiveness of our approach is demonstrated throu
gh unsupervised 3D semantic correspondence and shape segmentation.
**********

Deep Multimodal Fusion by Channel  Exchanging
Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, Junzhou Huang
Deep multimodal fusion by using multiple sources of data for classification or r
egression has exhibited a clear advantage over the unimodal counterpart on vario
us applications. Yet, current methods including aggregation-based and alignment-
based fusion are still inadequate in balancing the trade-off between inter-modal
 fusion and intra-modal processing, incurring a bottleneck of performance improv
ement. To this end, this paper proposes Channel-Exchanging-Network (CEN), a para
meter-free multimodal fusion framework that dynamically exchanges channels betwe
en sub-networks of different modalities. Specifically, the channel exchanging pr
ocess is self-guided by individual channel importance that is measured by the ma
gnitude of Batch-Normalization (BN) scaling factor during training. The validity
 of such exchanging process is also guaranteed by sharing convolutional filters
yet keeping separate BN layers across modalities, which, as an add-on benefit, a
llows our multimodal architecture to be almost as compact as a unimodal network.
 Extensive experiments on semantic segmentation via RGB-D data and image transla
tion through multi-domain input verify the effectiveness of our CEN compared to
current state-of-the-art methods. Detailed ablation studies have also been carri
ed out, which provably affirm the advantage of each component we propose. Our co
de is available at https://github.com/yikaiw/CEN.
**********

Hierarchically Organized Latent Modules for Exploratory Search in Morphogenetic
Systems
Mayalen Etcheverry, Clément Moulin-Frier, Pierre-Yves Oudeyer
Self-organization of complex morphological patterns from local interactions is a
fascinating phenomenon in many natural and artificial systems. In the artificial
world, typical examples of such morphogenetic systems are cellular automata. Yet
,
their mechanisms are often very hard to grasp and so far scientific discoveries
of
novel patterns have primarily been relying on manual tuning and ad hoc explorato

ry
search. The problem of automated diversity-driven discovery in these systems was
recently introduced [26, 62], highlighting that two key ingredients are autonomo
us
exploration and unsupervised representation learning to describe “relevant” degr
ees
of variations in the patterns. In this paper, we motivate the need for what we c
all
Meta-diversity search, arguing that there is not a unique ground truth interesti
ng
diversity as it strongly depends on the final observer and its motives. Using a
continuous game-of-life system for experiments, we provide empirical evidences
that relying on monolithic architectures for the behavioral embedding design ten
ds
to bias the final discoveries (both for hand-defined and unsupervisedly-learned
features) which are unlikely to be aligned with the interest of a final end-user
. To
address these issues, we introduce a novel dynamic and modular architecture that
enables unsupervised learning of a hierarchy of diverse representations. Combine
d
with intrinsically motivated goal exploration algorithms, we show that this syst
em
forms a discovery assistant that can efficiently adapt its diversity search towa
rds
preferences of a user using only a very small amount of user feedback.
**********
AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity
Silviu-Marian Udrescu, Andrew Tan, Jiahai Feng, Orisvaldo Neto, Tailin Wu, Max T
egmark
We present an improved method for symbolic regression that seeks to fit data to
formulas that are Pareto-optimal, in the sense of having the best accuracy for a
 given complexity. It improves on the previous state-of-the-art by typically bei
ng orders of magnitude more robust toward noise and bad data, and also by discov
ering many formulas that stumped previous methods. We develop a method for disco
vering generalized symmetries (arbitrary modularity in the computational graph o
f a formula) from gradient properties of a neural network fit. We use normalizin
g flows to generalize our symbolic regression method to probability distribution
s from which we only have samples, and employ statistical hypothesis testing to
accelerate robust brute-force search.
**********
Delay and Cooperation in Nonstochastic Linear Bandits
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.
**********
Probabilistic Orientation Estimation with Matrix Fisher Distributions
David Mohlin, Josephine Sullivan, Gérald Bianchi
This paper focuses on estimating probability distributions over the set of 3D ro
-
tations (SO(3)) using deep neural networks. Learning to regress models to the
set of rotations is inherently difficult due to differences in topology between
R^N and SO(3). We overcome this issue by using a neural network to out-
put the parameters for a matrix Fisher distribution since these parameters are
homeomorphic to R^9 . By using a negative log likelihood loss for this distri-
bution we get a loss which is convex with respect to the network outputs. By
optimizing this loss we improve state-of-the-art on several challenging applica-
ble datasets, namely Pascal3D+, ModelNet10-SO(3). Our code is available at

https://github.com/Davmo049/Publicproborientationestimationwithmatrix_fisherdistributions
**********
Minimax Dynamics of Optimally Balanced Spiking Networks of Excitatory and Inhibitory Neurons
Qianyi Li, Cengiz Pehlevan
Excitation-inhibition balance is ubiquitously observed in the cortex. Recent studies suggest an intriguing link between balance on fast timescales, tight balance, and efficient information coding with spikes. We further this connection by taking a principled approach to optimal balanced networks of excitatory (E) and inhibitory(I) neurons.  By deriving E-I spiking neural networks from greedy spike-based optimizations of constrained minimax objectives, we show that tight balance arises from correcting for deviations from the minimax optimum.  We predict specific neuron firing rates in the networks by solving the minimax problems, going beyond statistical theories of balanced networks. We design minimax objectives for reconstruction of an input signal, associative memory, and storage of manifold attractors, and derive from them E-I networks that perform the computation.
  Overall, we present a novel normative modeling approach for spiking E-I networks, going beyond the widely-used energy-minimizing networks that violate Dale’s law. Our networks can be used to model cortical circuits and computations.
**********
Telescoping Density-Ratio Estimation
Benjamin Rhodes, Kai Xu, Michael U. Gutmann
Density-ratio estimation via classification is a cornerstone of unsupervised learning. It has provided the foundation for state-of-the-art methods in representation learning and generative modelling, with the number of use-cases continuing to proliferate. However, it suffers from a critical limitation: it fails to accurately estimate ratios p/q for which the two densities differ significantly. Empirically, we find this occurs whenever the KL divergence between p and q exceeds tens of nats. To resolve this limitation, we introduce a new framework, telescoping density-ratio estimation (TRE), that enables the estimation of ratios between highly dissimilar densities in high-dimensional spaces. Our experiments demonstrate that TRE can yield substantial improvements over existing single-ratio methods for mutual information estimation, representation learning and energy-based modelling.
**********
Towards Deeper Graph Neural Networks with Differentiable Group Normalization
Kaixiong Zhou, Xiao Huang, Yuening Li, Daochen Zha, Rui Chen, Xia Hu
Graph neural networks (GNNs), which learn the representation of a node by aggregating its neighbors, have become an effective computational tool in downstream applications. Over-smoothing is one of the key issues which limit the performance of GNNs as the number of layers increases. It is because the stacked aggregators would make node representations converge to indistinguishable vectors. Several attempts have been made to tackle the issue by bringing linked node pairs close and unlinked pairs distinct. However, they often ignore the intrinsic community structures and would result in sub-optimal performance. The representations of nodes within the same community/class need be similar to facilitate the classification, while different classes are expected to be separated in embedding space. To bridge the gap, we introduce two over-smoothing metrics and a novel technique, i.e., differentiable group normalization (DGN). It normalizes nodes within the same group independently to increase their smoothness, and separates node distributions among different groups to significantly alleviate the over-smoothing issue. Experiments on real-world datasets demonstrate that DGN makes GNN models more robust to over-smoothing and achieves better performance with deeper GNNs.
**********
Stochastic Optimization for Performative Prediction
Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, Moritz Hardt
Requests for name changes in the electronic proceedings will be accepted with 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 Differentiable Programs with Admissible Neural Heuristics
Ameesh Shah, Eric Zhan, Jennifer Sun, Abhinav Verma, Yisong Yue, Swarat Chaudhuri
We study the problem of learning differentiable functions expressed as programs in a domain-specific language. Such programmatic models can offer benefits such as composability and interpretability; however, learning them requires optimizing over a combinatorial space of program "architectures". We frame this optimization problem as a search in a weighted graph whose paths encode top-down derivations of program syntax. Our key innovation is to view various classes of neural networks as continuous relaxations over the space of programs, which can then be used to complete any partial program. All the parameters of this relaxed program can be trained end-to-end, and the resulting training loss is an approximately admissible heuristic that can guide the combinatorial search. We instantiate our approach on top of the A* and Iterative Deepening Depth-First Search algorithms and use these algorithms to learn programmatic classifiers in three sequence classification tasks. Our experiments show that the algorithms outperform state-of-the-art methods for program learning, and that they discover programmatic classifiers that yield natural interpretations and achieve competitive accuracy.
**********
Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nystrom method
Michal Derezinski, Rajiv Khanna, Michael W. Mahoney
The Column Subset Selection Problem (CSSP) and the Nystrom method
are among the leading tools for constructing small low-rank
approximations of large datasets in machine learning and scientific
computing. A fundamental question in this area is: how well can a data subset of size k compete with the best rank k approximation?
We develop techniques which exploit spectral properties of the data
matrix to obtain improved approximation guarantees which go beyond the
standard worst-case analysis.
Our approach leads to significantly better bounds for datasets with
known rates of singular value decay, e.g., polynomial or exponential decay.
Our analysis also reveals an intriguing phenomenon: the approximation
factor as a function of k may exhibit multiple peaks and valleys,
which we call a multiple-descent curve.
A lower bound we establish shows that this behavior is not an artifact
of our analysis, but rather it is an inherent property of the CSSP and
Nystrom tasks. Finally, using the example of a radial basis function (RBF)
kernel, we show that both our improved bounds and the multiple-descent
curve can be observed on real datasets simply by varying the RBF parameter.
**********
Domain Adaptation as a Problem of Inference on Graphical Models
Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, QINGSONG LIU, Clark Glymour
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Network size and size of the weights in memorization with two-layers neural networks
Sebastien Bubeck, Ronen Eldan, Yin Tat Lee, Dan Mikulincer
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Certifying Strategyproof Auction Networks

Michael Curry, Ping-yeh Chiang, Tom Goldstein, John Dickerson
Optimal auctions maximize a seller's expected revenue subject to individual rati onality and strategyproofness for the buyers.  Myerson's seminal work in 1981 se ttled the case of auctioning a single item; however, subsequent decades of work have yielded little progress moving beyond a single item, leaving the design of revenue-maximizing auctions as a central open problem in the field of mechanism design. A recent thread of work in ``differentiable economics'' has used tools f rom modern deep learning to instead learn good mechanisms. We focus on the Regre tNet architecture, which can represent auctions with arbitrary numbers of items and participants; it is trained to be empirically strategyproof, but the propert y is never exactly verified leaving potential loopholes for market participants to exploit. We propose ways to explicitly verify strategyproofness under a parti cular valuation profile using techniques from the neural network verification li terature. Doing so requires making several modifications to the RegretNet archit ecture in order to represent it exactly in an integer program. We train our netw ork and produce certificates in several settings, including settings for which t he optimal strategyproof mechanism is not known.
**********

Continual Learning of Control Primitives : Skill Discovery via Reset-Games
Kelvin Xu, Siddharth Verma, Chelsea Finn, Sergey Levine
Reinforcement learning has the potential to automate the acquisition of behavior  in complex settings, but in order for it to be successfully deployed, a number of practical challenges must be addressed. First, in real world settings, when a n agent attempts a tasks and fails, the environment must somehow "reset" so that  the agent can attempt the task again. While easy in simulation, this could requ ire considerable human effort in the real world, especially if the number of tri als is very large. Second, real world learning is often limited by challenges in  exploration, as complex, temporally extended behavior is often times difficult to acquire with random exploration. In this work, we show how a single method ca n allow an agent to acquire skills with minimal supervision while removing the n eed for resets. We do this by exploiting the insight that the need to reset" an agent to a broad set of initial states for a learning task provides a natural se tting to learn a diverse set ofreset-skills." We propose a general-sum game form ulation that naturally balances the objective of resetting and learning skills, and demonstrate that this approach improves performance on reset-free tasks, and  additionally show that the skills we obtain can be used to significantly accele rate downstream learning.
**********

HOI Analysis: Integrating and Decomposing Human-Object Interaction
Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Cewu Lu
Human-Object Interaction (HOI) consists of human, object and implicit interactio n/verb. Different from previous methods that directly map pixels to HOI semantic s, we propose a novel perspective for HOI learning in an analytical manner. In a nalogy to Harmonic Analysis, whose goal is to study how to represent the signals  with the superposition of basic waves, we propose the HOI Analysis. We argue th at coherent HOI can be decomposed into isolated human and object. Meanwhile, iso lated human and object can also be integrated into coherent HOI again. Moreover,  transformations between human-object pairs with the same HOI can also be easier  approached with integration and decomposition. As a result, the implicit verb w ill be represented in the transformation function space. In light of this, we pr opose an Integration-Decomposition Network (IDN) to implement the above transfor mations and achieve state-of-the-art performance on widely-used HOI detection be nchmarks. Code is available at https://github.com/DirtyHarryLYL/HAKE-Action-Torc h/tree/IDN-(Integrating-Decomposing-Network).
**********

Strongly local p-norm-cut algorithms for semi-supervised learning and local grap h clustering
Meng Liu, David F. Gleich
Graph based semi-supervised learning is the problem of learning a labeling funct ion for the graph nodes given a few example nodes, often called seeds, usually u

nder the assumption that the graph’s edges indicate similarity of labels.
This is closely related to the local graph clustering or community detection pro blem of finding a cluster or community of nodes around a given seed.
For this problem, we propose a novel generalization of random walk, diffusion, o r smooth function methods in the literature to a convex p-norm cut function.
The need for our p-norm methods is that, in our study of existing methods, we fi nd those principled methods based on eigenvector, spectral, random walk, or line ar system often have difficulty capturing the correct boundary of a target label or target cluster.
In contrast, 1-norm or maxflow-mincut based methods capture the boundary, but ca nnot grow from small seed set; hybrid procedures that use both have many hard to set parameters.
In this paper, we propose a generalization of the objective function behind thes e methods involving p-norms.
To solve the p-norm cut problem we give a strongly local algorithm -- one whose runtime depends on the size of the output rather than the size of the graph.
Our method can be thought as a nonlinear generalization of the Anderson-Chung-La ng push procedure to approximate a personalized PageRank vector efficiently.
Our procedure is general and can solve other types of nonlinear objective functi ons, such as p-norm variants of Huber losses. We provide a theoretical analysis of finding planted target clusters with our method and show that the p-norm cut functions improve on the standard Cheeger inequalities for random walk and spect ral methods. Finally, we demonstrate the speed and accuracy of our new method in synthetic and real world datasets.
**********

Deep Direct Likelihood Knockoffs
Mukund Sudarshan, Wesley Tansey, Rajesh Ranganath
Predictive modeling often uses black box machine learning methods, such as deep neural networks, to achieve state-of-the-art performance. In scientific domains, the scientist often wishes to discover which features are actually important fo r making the predictions. These discoveries may lead to costly follow-up experim ents and as such it is important that the error rate on discoveries is not too h igh. Model-X knockoffs enable important features to be discovered with control o f the false discovery rate (FDR). However, knockoffs require rich generative mod els capable of accurately modeling the knockoff features while ensuring they obe y the so-called "swap" property. We develop Deep Direct Likelihood Knockoffs (DD LK), which directly minimizes the KL divergence implied by the knockoff swap pro perty. DDLK consists of two stages: it first maximizes the explicit likelihood o f the features, then minimizes the KL divergence between the joint distribution of features and knockoffs and any swap between them. To ensure that the generate d knockoffs are valid under any possible swap, DDLK uses the Gumbel-Softmax tric k to optimize the knockoff generator under the worst-case swap. We find DDLK has higher power than baselines while controlling the false discovery rate on a var iety of synthetic and real benchmarks including a task involving the largest COV ID-19 health record dataset in the United States.
**********

Meta-Neighborhoods
Siyuan Shan, Yang Li, Junier B. Oliva
Making an adaptive prediction based on input is an important ability for general artificial intelligence. In this work, we step forward in this direction and pr opose a semi-parametric method, Meta-Neighborhoods, where predictions are made a daptively to the neighborhood of the input. We show that Meta-Neighborhoods is a generalization of k-nearest-neighbors. Due to the simpler manifold structure ar ound a local neighborhood, Meta-Neighborhoods represent the predictive distribut ion p(y | x) more accurately. To reduce memory and computation overheads, we pro pose induced neighborhoods that summarize the training data into a much smaller dictionary. A meta-learning based training mechanism is then exploited to jointl y learn the induced neighborhoods and the model. Extensive studies demonstrate t he superiority of our method.
**********

Neural Dynamic Policies for End-to-End Sensorimotor Learning
Shikhar Bahl, Mustafa Mukadam, Abhinav Gupta, Deepak Pathak
The current dominant paradigm in sensorimotor control, whether imitation or reinforcement learning, is to train policies directly in raw action spaces such as torque, joint angle, or end-effector position. This forces the agent to make decision at each point in training, and hence, limits the scalability to continuous, high-dimensional, and long-horizon tasks. In contrast, research in classical robotics has, for a long time, exploited dynamical systems as a policy representation to learn robot behaviors via demonstrations. These techniques, however, lack the flexibility and generalizability provided by deep learning or deep reinforcement learning and have remained under-explored in such settings. In this work, we begin to close this gap and embed dynamics structure into deep neural network-based policies by reparameterizing action spaces with differential equations. We propose Neural Dynamic Policies (NPDs) that make predictions in trajectory distribution space as opposed to prior policy learning methods where action represents the raw control space. The embedded structure allows us to perform end-to-end policy learning under both reinforcement and imitation learning setups. We show that NDPs achieve better or comparable performance to state-of-the-art approaches on many robotic control tasks using both reward-based training and demonstrations. Project video and code are available at: https://shikharbahl.github.io/neural-dynamic-policies/.
**********
A new inference approach for training shallow and deep generalized linear models of noisy interacting neurons
Gabriel Mahuas, Giulio Isacchini, Olivier Marre, Ulisse Ferrari, Thierry Mora
Generalized linear models are one of the most efficient paradigms for predicting the correlated stochastic activity of neuronal networks in response to external stimuli, with applications in many brain areas. However, when dealing with complex stimuli, the inferred coupling parameters often do not generalise across different stimulus statistics, leading to degraded performance and blowup instabilities. Here, we develop a two-step inference strategy that allows us to train robust generalised linear models of interacting neurons, by explicitly separating the effects of correlations in the stimulus from network interactions in each training step. Applying this approach to the responses of retinal ganglion cells to complex visual stimuli, we show that, compared to classical methods, the models trained in this way exhibit improved performance, are more stable, yield robust interaction networks, and generalise well across complex visual statistics. The method can be extended to deep convolutional neural networks, leading to models with high predictive accuracy for both the neuron firing rates and their correlations.
**********
Decision-Making with Auto-Encoding Variational Bayes
Romain Lopez, Pierre Boyeau, Nir Yosef, Michael Jordan, Jeffrey Regier
To make decisions based on a model fit with auto-encoding variational Bayes (AEVB), practitioners often let the variational distribution serve as a surrogate for the posterior distribution. This approach yields biased estimates of the expected risk, and therefore leads to poor decisions for two reasons. First, the model fit with AEVB may not equal the underlying data distribution. Second, the variational distribution may not equal the posterior distribution under the fitted model.
We explore how fitting the variational distribution based on several objective functions other than the ELBO, while continuing to fit the generative model based on the ELBO, affects the quality of downstream decisions.
For the probabilistic principal component analysis model, we investigate how importance sampling error, as well as the bias of the model parameter estimates, varies across several approximate posteriors when used as proposal distributions.
Our theoretical results suggest that a posterior approximation distinct from the variational distribution should be used for making decisions. Motivated by these theoretical results, we propose learning several approximate proposals for the best model and combining them using multiple importance sampling for decision-m

aking. In addition to toy examples, we present a full-fledged case study of sing le-cell RNA sequencing. In this challenging instance of multiple hypothesis test ing, our proposed approach surpasses the current state of the art.
**********

Attribution Preservation in Network Compression for Reliable Network Interpretat ion

Geondo Park, June Yong Yang, Sung Ju Hwang, Eunho Yang

Neural networks embedded in safety-sensitive applications such as self-driving c ars and wearable health monitors rely on two important techniques: input attribu tion for hindsight analysis and network compression to reduce its size for edge- computing. In this paper, we show that these seemingly unrelated techniques conf lict with each other as network compression deforms the produced attributions, w hich could lead to dire consequences for mission-critical applications. This phe nomenon arises due to the fact that conventional network compression methods onl y preserve the predictions of the network while ignoring the quality of the attr ibutions. To combat the attribution inconsistency problem, we present a framewor k that can preserve the attributions while compressing a network. By employing t he Weighted Collapsed Attribution Matching regularizer, we match the attribution maps of the network being compressed to its pre-compression former self. We dem onstrate the effectiveness of our algorithm both quantitatively and qualitativel y on diverse compression methods.
**********

Feature Importance Ranking for Deep Learning

Maksymilian Wojtas, Ke Chen

Feature importance ranking has become a powerful tool for explainable AI. Howeve r, its nature of combinatorial optimization poses a great challenge for deep lea rning. In this paper, we propose a novel dual-net architecture consisting of ope rator and selector for discovery of an optimal feature subset of a fixed size an d ranking the importance of those features in the optimal subset simultaneously. During learning, the operator is trained for a supervised learning task via opt imal feature subset candidates generated by the selector that learns predicting the learning performance of the operator working on different optimal subset can didates. We develop an alternate learning algorithm that trains two nets jointly and incorporates a stochastic local search procedure into learning to address t he combinatorial optimization challenge. In deployment, the selector generates a n optimal feature subset and ranks feature importance, while the operator makes predictions based on the optimal subset for test data. A thorough evaluation on synthetic, benchmark and real data sets suggests that our approach outperforms s everal state-of-the-art feature importance ranking and supervised feature select ion methods. (Our source code is available: https://github.com/maksym33/FeatureI mportanceDL)
**********

Causal Estimation with Functional Confounders

Aahlad Puli, Adler Perotte, Rajesh Ranganath

Causal inference relies on two fundamental assumptions: ignorability and positiv ity. We study causal inference when the true confounder value can be expressed a s a function of the observed data; we call this setting estimation with function al confounders (EFC). In this setting ignorability is satisfied, however positiv ity is violated, and causal inference is impossible in general. We consider two scenarios where causal effects are estimable. First, we discuss interventions on a part of the treatment called functional interventions and a sufficient condit ion for effect estimation of these interventions called functional positivity. S econd, we develop conditions for nonparametric effect estimation based on the gr adient fields of the functional confounder and the true outcome function. To est imate effects under these conditions, we develop Level-set Orthogonal Descent Es timation (LODE). Further, we prove error bounds on LODE’s effect estimates, eval uate our methods on simulated and real data, and empirically demonstrate the val ue of EFC.
**********

Model Inversion Networks for Model-Based Optimization

Aviral Kumar, Sergey Levine
This work addresses data-driven optimization problems, where the goal is to find an input that maximizes an unknown score or reward function given access to a dataset of inputs with corresponding scores. When the inputs are high-dimensional and valid inputs constitute a small subset of this space (e.g., valid protein sequences or valid natural images), such model-based optimization problems become exceptionally difficult, since the optimizer must avoid out-of-distribution and invalid inputs. We propose to address such problems with model inversion networks (MINs), which learn an inverse mapping from scores to inputs. MINs can scale to high-dimensional input spaces and leverage offline logged data for both contextual and non-contextual optimization problems. MINs can also handle both purely offline data sources and active data collection. We evaluate MINs on high- dimensional model-based optimization problems over images, protein designs, and neural network controller parameters, and bandit optimization from logged data.
**********
Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks
Umut Simsekli, Ozan Sener, George Deligiannidis, Murat A. Erdogdu
Despite its success in a wide range of applications, characterizing the generalization properties of stochastic gradient descent (SGD) in non-convex deep learning problems is still an important challenge. While modeling the trajectories of SGD via stochastic differential equations (SDE) under heavy-tailed gradient noise has recently shed light over several peculiar characteristics of SGD, a rigorous treatment of the generalization properties of such SDEs in a learning theoretical framework is still missing. Aiming to bridge this gap, in this paper, we prove generalization bounds for SGD under the assumption that its trajectories can be well-approximated by a \emph{Feller process}, which defines a rich class of Markov processes that include several recent SDE representations (both Brownian or heavy-tailed) as its special case. We show that the generalization error can be controlled by the \emph{Hausdorff dimension} of the trajectories, which is intimately linked to the tail behavior of the driving process. Our results imply that heavier-tailed processes should achieve better generalization; hence, the tail-index of the process can be used as a notion of ``capacity metric''. We support our theory with experiments on deep neural networks illustrating that the proposed capacity metric accurately estimates the generalization error, and it does not necessarily grow with the number of parameters unlike the existing capacity metrics in the literature.
**********
Exact expressions for double descent and implicit regularization via surrogate random design
Michal Derezinski, Feynman T. Liang, Michael W. Mahoney
Double descent refers to the phase transition that is exhibited by
the generalization error of unregularized learning models when varying the ratio
between the number of parameters and the number of training
samples. The recent success of highly over-parameterized machine learning
models such as deep neural networks has motivated a theoretical analysis of
the double descent phenomenon in classical models such as linear
regression which can also generalize well in the over-parameterized
regime. We provide the first exact non-asymptotic
expressions for double descent of the minimum norm linear
estimator. Our approach involves constructing a special
determinantal point process which we call surrogate random
design, to replace the standard i.i.d. design of the training
sample. This surrogate design admits exact expressions for the mean
squared error of the estimator while preserving the key properties
of the standard design. We also establish an exact implicit
regularization result for over-parameterized training samples. In
particular, we show that, for the surrogate design, the implicit bias
of the unregularized minimum norm estimator precisely corresponds to
solving a ridge-regularized least squares problem on the population
distribution. In our analysis we introduce a new mathematical tool of

independent interest: the class of random matrices for which
determinant commutes with expectation.
**********

Certifying Confidence via Randomized Smoothing
Aounon Kumar, Alexander Levine, Soheil Feizi, Tom Goldstein
Randomized smoothing has been shown to provide good certified-robustness guarant
ees for high-dimensional classification problems.
It uses the probabilities of predicting the top two most-likely classes around a
n input point under a smoothing distribution to generate a certified radius for
a classifier's prediction.
However, most smoothing methods do not give us any information about the \emph{c
onfidence} with which the underlying classifier (e.g., deep neural network) make
s a prediction.
In this work, we propose a method to generate certified radii for the prediction
 confidence of the smoothed classifier.
We consider two notions for quantifying confidence: average prediction score of
a class and the margin by which the average prediction score of one class exceed
s that of another.
We modify the Neyman-Pearson lemma (a key theorem in randomized smoothing) to de
sign a procedure for computing the certified radius where the confidence is guar
anteed to stay above a certain threshold.
Our experimental results on CIFAR-10 and ImageNet datasets show that using infor
mation about the distribution of the confidence scores allows us to achieve a si
gnificantly better certified radius than ignoring it.
Thus, we demonstrate that extra information about the base classifier at the inp
ut point can help improve certified guarantees for the smoothed classifier.
Code for the experiments is available at \url{https://github.com/aounon/cdf-smoo
thing}.
**********

Learning Physical Constraints with Neural Projections
Shuqi Yang, Xingzhe He, Bo Zhu
We propose a new family of neural networks to predict the behaviors of physical
systems by learning their underpinning constraints. A neural projection operator
 lies at the heart of our approach, composed of a lightweight network with an em
bedded recursive architecture that interactively enforces learned underpinning c
onstraints and predicts the various governed behaviors of different physical sys
tems. Our neural projection operator is motivated by the position-based dynamics
 model that has been used widely in game and visual effects industries to unify
the various fast physics simulators. Our method can automatically and effectivel
y uncover a broad range of constraints from observation point data, such as leng
th, angle, bending, collision, boundary effects, and their arbitrary combination
s, without any connectivity priors. We provide a multi-group point representatio
n in conjunction with a configurable network connection mechanism to incorporate
 prior inputs for processing complex physical systems. We demonstrated the effic
acy of our approach by learning a set of challenging physical systems all in a u
nified and simple fashion including: rigid bodies with complex geometries, ropes
 with varying length and bending, articulated soft and rigid bodies, and multi-o
bject collisions with complex boundaries.
**********

Robust Optimization for Fairness with Noisy Protected Groups
Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, Mic
hael Jordan
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Noise-Contrastive Estimation for Multivariate Point Processes
Hongyuan Mei, Tom Wan, Jason Eisner
The log-likelihood of a generative model often involves both positive and negati

ve terms. For a temporal multivariate point process, the negative term sums over all the possible event types at each time and also integrates over all the possible times. As a result, maximum likelihood estimation is expensive. We show how to instead apply a version of noise-contrastive estimation---a general parameter estimation method with a less expensive stochastic objective. Our specific instantiation of this general idea works out in an interestingly non-trivial way and has provable guarantees for its optimality, consistency and efficiency. On several synthetic and real-world datasets, our method shows benefits: for the model to achieve the same level of log-likelihood on held-out data, our method needs considerably fewer function evaluations and less wall-clock time.
**********

A Game-Theoretic Analysis of the Empirical Revenue Maximization Algorithm with Endogenous Sampling
Xiaotie Deng, Ron Lavi, Tao Lin, Qi Qi, Wenwei WANG, Xiang Yan
The Empirical Revenue Maximization (ERM) is one of the most important price learning algorithms in auction design: as the literature shows it can learn approximately optimal reserve prices for revenue-maximizing auctioneers in both repeated auctions and uniform-price auctions. However, in these applications the agents who provide inputs to ERM have incentives to manipulate the inputs to lower the outputted price. We generalize the definition of an incentive-awareness measure proposed by Lavi et al (2019), to quantify the reduction of ERM's outputted price due to a change of m>=1 out of N input samples, and provide specific convergence rates of this measure to zero as N goes to infinity for different types of input distributions. By adopting this measure, we construct an efficient, approximately incentive-compatible, and revenue-optimal learning algorithm using ERM in repeated auctions against non-myopic bidders, and show approximate group incentive-compatibility in uniform-price auctions.
**********

Neural Path Features and Neural Path Kernel : Understanding the role of gates in deep learning
Chandrashekar Lakshminarayanan, Amit Vikram Singh
Rectified linear unit (ReLU) activations can also be thought of as 'gates', which, either pass or stop their pre-activation input when they are 'on' (when the pre-activation input is positive) or 'off' (when the pre-activation input is negative) respectively. A deep neural network (DNN) with ReLU activations has many gates, and the on/off status of each gate changes across input examples as well as network weights. For a given input example, only a subset of gates are 'active', i.e., on, and the sub-network of weights connected to these active gates is responsible for producing the output. At randomised initialisation, the active sub-network corresponding to a given input example is random. During training, as the weights are learnt, the active sub-networks are also learnt, and could hold valuable information.
**********

Multiscale Deep Equilibrium Models
Shaojie Bai, Vladlen Koltun, J. Zico Kolter
We propose a new class of implicit networks, the multiscale deep equilibrium model (MDEQ), suited to large-scale and highly hierarchical pattern recognition domains. An MDEQ directly solves for and backpropagates through the equilibrium points of multiple feature resolutions simultaneously, using implicit differentiation to avoid storing intermediate states (and thus requiring only O(1) memory consumption). These simultaneously-learned multi-resolution features allow us to train a single model on a diverse set of tasks and loss functions, such as using a single MDEQ to perform both image classification and semantic segmentation. We illustrate the effectiveness of this approach on two large-scale vision tasks: ImageNet classification and semantic segmentation on high-resolution images from the Cityscapes dataset. In both settings, MDEQs are able to match or exceed the performance of recent competitive computer vision models: the first time such performance and scale have been achieved by an implicit deep learning approach. The code and pre-trained models are at https://github.com/locuslab/mdeq.
**********

Sparse Graphical Memory for Robust Planning
Scott Emmons, Ajay Jain, Misha Laskin, Thanard Kurutach, Pieter Abbeel, Deepak Pathak
To operate effectively in the real world, agents should be able to act from high-dimensional raw sensory input such as images and achieve diverse goals across long time-horizons. Current deep reinforcement and imitation learning methods can learn directly from high-dimensional inputs but do not scale well to long-horizon tasks. In contrast, classical graphical methods like A* search are able to solve long-horizon tasks, but assume that the state space is abstracted away from raw sensory input. Recent works have attempted to combine the strengths of deep learning and classical planning; however, dominant methods in this domain are still quite brittle and scale poorly with the size of the environment. We introduce Sparse Graphical Memory (SGM), a new data structure that stores states and feasible transitions in a sparse memory. SGM aggregates states according to a novel two-way consistency objective, adapting classic state aggregation criteria to goal-conditioned RL: two states are redundant when they are interchangeable both as goals and as starting states. Theoretically, we prove that merging nodes according to two-way consistency leads to an increase in shortest path lengths that scales only linearly with the merging threshold. Experimentally, we show that SGM significantly outperforms current state of the art methods on long horizon, sparse-reward visual navigation tasks. Project video and code are available at https://sites.google.com/view/sparse-graphical-memory.
**********
Second Order PAC-Bayesian Bounds for the Weighted Majority Vote
Andres Masegosa, Stephan Lorenzen, Christian Igel, Yevgeny Seldin
We present a novel analysis of the expected risk of weighted majority vote in multiclass classification. The analysis takes correlation of predictions by ensemble members into account and provides a bound that is amenable to efficient minimization, which yields improved weighting for the majority vote. We also provide a specialized version of our bound for binary classification, which allows to exploit additional unlabeled data for tighter risk estimation. In experiments, we apply the bound to improve weighting of trees in random forests and show that, in contrast to the commonly used first order bound, minimization of the new bound typically does not lead to degradation of the test error of the ensemble.
**********
Dirichlet Graph Variational Autoencoder
Jia Li, Jianwei Yu, Jiajin Li, Honglei Zhang, Kangfei Zhao, Yu Rong, Hong Cheng, Junzhou Huang
Graph Neural Networks (GNN) and Variational Autoencoders (VAEs) have been widely used in modeling and generating graphs with latent factors. However there is no clear explanation of what these latent factors are and why they perform well. In this work, we present Dirichlet Graph Variational Autoencoder (DGVAE) with graph cluster memberships as latent factors. Our study connects VAEs based graph generation and balanced graph cut, and provides a new way to understand and improve the internal mechanism of VAEs based graph generation. Specifically, we first interpret the reconstruction term of DGVAE as balanced graph cut in a principled way. Furthermore, motivated by the low pass characteristics in balanced graph cut, we propose a new variant of GNN named Heatts to encode the input graph into cluster memberships. Heatts utilizes the Taylor series for fast computation of Heat kernels and has better low pass characteristics than Graph Convolutional Networks (GCN). Through experiments on graph generation and graph clustering, we demonstrate the effectiveness of our proposed framework.
**********
Modeling Task Effects on Meaning Representation in the Brain via Zero-Shot MEG Prediction
Mariya Toneva, Otilia Stretcu, Barnabas Poczos, Leila Wehbe, Tom M. Mitchell
How meaning is represented in the brain is still one of the big open questions in neuroscience. Does a word (e.g., bird) always have the same representation, or does the task under which the word is processed alter its representation (answering can you eat it?" versuscan it fly?")? The brain activity of subjects who re

ad the same word while performing different semantic tasks has been shown to dif
fer across tasks. However, it is still not understood how the task itself contri
butes to this difference. In the current work, we study Magnetoencephalography (
MEG) brain recordings of participants tasked with answering questions about conc
rete nouns. We investigate the effect of the task (i.e. the question being asked
) on the processing of the concrete noun by predicting the millisecond-resolutio
n MEG recordings as a function of both the semantics of the noun and the task. U
sing this approach, we test several hypotheses about the task-stimulus interacti
ons by comparing the zero-shot predictions made by these hypotheses for novel ta
sks and nouns not seen during training. We find that incorporating the task sema
ntics significantly improves the prediction of MEG recordings, across participan
ts. The improvement occurs 475-550ms after the participants first see the word,
which corresponds to what is considered to be the ending time of semantic proces
sing for a word. These results suggest that only the end of semantic processing
of a word is task-dependent, and pose a challenge for future research to formula
te new hypotheses for earlier task effects as a function of the task and stimuli
.
**********

Counterfactual Vision-and-Language Navigation: Unravelling the Unseen
Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Javen Qinfeng Shi, Anton van den
Hengel
The task of vision-and-language navigation (VLN) requires an agent to follow tex
t instructions to find its way through simulated household environments. A promi
nent challenge is to train an agent capable of generalising to new environments
at test time, rather than one that simply memorises trajectories and visual deta
ils observed during training. We propose a new learning strategy that learns bot
h from observations and generated counterfactual environments. We describe an ef
fective algorithm to generate counterfactual observations on the fly for VLN, as
 linear combinations of existing environments. Simultaneously, we encourage the
agent's actions to remain stable between original and counterfactual environment
s through our novel training objective-effectively removing the spurious feature
s that otherwise bias the agent. Our experiments show that this technique provid
es significant improvements in generalisation on benchmarks for Room-to-Room nav
igation and Embodied Question Answering.
**********

Robust Quantization: One Model to Rule Them All
moran shkolnik, Brian Chmiel, Ron Banner, Gil Shomron, Yury Nahshan, Alex Bronst
ein, Uri Weiser
Neural network quantization methods often involve simulating the quantization pr
ocess during training, making the trained model highly dependent on the target b
it-width and precise way quantization is performed.  Robust quantization offers
an alternative approach with improved tolerance to different classes of data-typ
es and quantization policies. It opens up new exciting applications where the qu
antization process is not static and can vary to meet different circumstances an
d implementations. To address this issue, we propose a method that provides intr
insic robustness to the model against a broad range of quantization processes.
Our method is motivated by theoretical arguments and enables us to store a singl
e generic model capable of operating at various bit-widths and quantization poli
cies. We validate our method's effectiveness on different ImageNet Models. A ref
erence implementation accompanies the paper.
**********

Enabling certification of verification-agnostic networks via memory-efficient se
midefinite programming
Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, J
onathan Uesato, Rudy R. Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow,
 Percy S. Liang, Pushmeet Kohli
Requests for name changes in the electronic proceedings will be accepted with 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.

**********
Federated Accelerated Stochastic Gradient Descent
Honglin Yuan, Tengyu Ma
We propose Federated Accelerated Stochastic Gradient Descent (FedAc), a principled acceleration of Federated Averaging (FedAvg, also known as Local SGD) for distributed optimization. FedAc is the first provable acceleration of FedAvg that improves convergence speed and communication efficiency on various types of convex functions. For example, for strongly convex and smooth functions, when using M workers, the previous state-of-the-art FedAvg analysis can achieve a linear speedup in M if given M rounds of synchronization, whereas FedAc only requires M^■ rounds. Moreover, we prove stronger guarantees for FedAc when the objectives are third-order smooth. Our technique is based on a potential-based perturbed iterate analysis, a novel stability analysis of generalized accelerated SGD, and a strategic tradeoff between acceleration and stability.
**********
Robust Density Estimation under Besov IPM Losses
Ananya Uppal, Shashank Singh, Barnabas Poczos
We study minimax convergence rates of nonparametric density estimation under the Huber contamination model, in which a ``contaminated'' proportion of the data comes from an unknown outlier distribution. We provide the first results for this problem under a large family of losses, called Besov integral probability metrics (IPMs), that include L^p, Wasserstein, Kolmogorov-Smirnov, Cramer-von Mises, and other commonly used metrics. Under a range of smoothness assumptions on the population and outlier distributions, we show that a re-scaled thresholding wavelet estimator converges at the minimax optimal rate under a wide variety of losses and also exhibits optimal dependence on the contamination proportion. We also provide a purely data-dependent extension of the estimator that adapts to both an unknown contamination proportion and the unknown smoothness of the true density. Finally, based on connections shown recently between density estimation under IPM losses and generative adversarial networks (GANs), we show that certain GAN architectures are robustly minimax optimal.
**********
An analytic theory of shallow networks dynamics for hinge loss classification
Franco Pellegrini, Giulio Biroli
Neural networks have been shown to perform incredibly well in classification tasks over structured high-dimensional datasets. However, the learning dynamics of such networks is still poorly understood. In this paper we study in detail the training dynamics of a simple type of neural network: a single hidden layer trained to perform a classification task. We show that in a suitable mean-field limit this case maps to a single-node learning problem with a time-dependent dataset determined self-consistently from the average nodes population. We specialize our theory to the prototypical case of a linearly separable dataset and a linear hinge loss, for which the dynamics can be explicitly solved in the infinite dataset limit. This allow us to address in a simple setting several phenomena appearing in modern networks such as slowing down of training dynamics, crossover between feature and lazy learning, and overfitting. Finally, we asses the limitations of mean-field theory by studying the case of large but finite number of nodes and of training samples.
**********
Fixed-Support Wasserstein Barycenters: Computational Hardness and Fast Algorithm
Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, Michael Jordan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Learning to Orient Surfaces by Self-supervised Spherical CNNs
Riccardo Spezialetti, Federico Stella, Marlon Marcon, Luciano Silva, Samuele Salti, Luigi Di Stefano
Defining and reliably finding a canonical orientation for 3D surfaces is key to

many Computer Vision and Robotics applications. This task is commonly addressed by handcrafted algorithms exploiting geometric cues deemed as distinctive and robust by the designer. Yet, one might conjecture that humans learn the notion of the inherent orientation of 3D objects from experience and that machines may do so alike. In this work, we show the feasibility of learning a robust canonical orientation for surfaces represented as point clouds. Based on the observation that the quintessential property of a canonical orientation is equivariance to 3D rotations, we propose to employ Spherical CNNs, a recently introduced machinery that can learn equivariant representations defined on the Special Ortoghonal group SO(3). Specifically, spherical correlations compute feature maps whose elements define 3D rotations. Our method learns such feature maps from raw data by a self-supervised training procedure and robustly selects a rotation to transform the input point cloud into a learned canonical orientation. Thereby, we realize the first end-to-end learning approach to define and extract the canonical orientation of 3D shapes, which we aptly dub Compass. Experiments on several public datasets prove its effectiveness at orienting local surface patches as well as whole objects.
**********

Adam with Bandit Sampling for Deep Learning
Rui Liu, Tianyi Wu, Barzan Mozafari
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Parabolic Approximation Line Search for DNNs
Maximus Mutschler, Andreas Zell
A major challenge in current optimization research for deep learning is to automatically find optimal step sizes for each update step. The optimal step size is closely related to the shape of the loss in the update step direction. However, this shape has not yet been examined in detail. This work shows empirically that the sample loss over lines in negative gradient direction is mostly convex and well suited for one-dimensional parabolic approximations. Exploiting this parabolic property we introduce a simple and robust line search approach, which performs loss-shape dependent update steps. Our approach combines well-known methods such as parabolic approximation, line search and conjugate gradient, to perform efficiently. It successfully competes with common and state-of-the-art optimization methods on a large variety of experiments without the need of hand-designed step size schedules. Thus, it is of interest for objectives where step-size schedules are unknown or do not perform well. Our excessive evaluation includes multiple comprehensive hyperparameter grid searches on several datasets and architectures. We provide proof of convergence for an adapted scenario. Finally, we give a general investigation of exact line searches in the context of sample losses and exact losses, including their relation to our line search approach.
**********

Agnostic Learning of a Single Neuron with Gradient Descent
Spencer Frei, 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.
**********

Statistical Efficiency of Thompson Sampling for Combinatorial Semi-Bandits
Pierre Perrault, Etienne Boursier, Michal Valko, Vianney Perchet
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Analytic Characterization of the Hessian in Shallow ReLU Models: A Tale of Symme

try
Yossi Arjevani, Michael Field
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Generative causal explanations of black-box classifiers
Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, Mark Davenport
We develop a method for generating causal post-hoc explanations of black-box classifiers based on a learned low-dimensional representation of the data. The explanation is causal in the sense that changing learned latent factors produces a change in the classifier output statistics. To construct these explanations, we design a learning framework that leverages a generative model and information-theoretic measures of causal influence. Our objective function encourages both the generative model to faithfully represent the data distribution and the latent factors to have a large causal influence on the classifier output. Our method learns both global and local explanations, is compatible with any classifier that admits class probabilities and a gradient, and does not require labeled attributes or knowledge of causal structure. Using carefully controlled test cases, we provide intuition that illuminates the function of our causal objective. We then demonstrate the practical utility of our method on image recognition tasks.
**********

Sub-sampling for Efficient Non-Parametric Bandit Exploration
Dorian Baudry, Emilie Kaufmann, Odalric-Ambrym Maillard
In this paper we propose the first multi-armed bandit algorithm based on re-sampling that achieves asymptotically optimal regret simultaneously for different families of arms (namely Bernoulli, Gaussian and Poisson distributions). Unlike Thompson Sampling which requires to specify a different prior to be optimal in each case, our proposal RB-SDA does not need any distribution-dependent tuning. RB-SDA belongs to the family of Sub-sampling Duelling Algorithms (SDA) which combines the sub-sampling idea first used by the BESA and SSMC algorithms with different sub-sampling schemes. In particular, RB-SDA uses Random Block sampling. We perform an experimental study assessing the flexibility and robustness of this promising novel approach for exploration in bandit models.
**********

Learning under Model Misspecification: Applications to Variational and Ensemble methods
Andres Masegosa
Virtually any model we use in machine learning to make predictions does not perfectly represent reality. So, most of the learning happens under model misspecification. In this work, we present a novel analysis of the generalization performance of Bayesian model averaging under model misspecification and i.i.d. data using a new family of second-order PAC-Bayes bounds. This analysis shows, in simple and intuitive terms, that Bayesian model averaging provides suboptimal generalization performance when the model is misspecified. In consequence, we provide strong theoretical arguments showing that Bayesian methods are not optimal for learning predictive models, unless the model class is perfectly specified. Using novel second-order PAC-Bayes bounds, we derive a new family of Bayesian-like algorithms, which can be implemented as variational and ensemble methods. The output of these algorithms is a new posterior distribution, different from the Bayesian posterior, which induces a posterior predictive distribution with better generalization performance. Experiments with Bayesian neural networks illustrate these findings.
**********

Language Through a Prism: A Spectral Approach for Multiscale Language Representations
Alex Tamkin, Dan Jurafsky, Noah Goodman
Language exhibits structure at a wide range of scales, from subwords to words, s

entences, paragraphs, and documents. We propose building models that isolate scale-specific information in deep representations, and develop methods for encouraging models during training to learn more about particular scales of interest. Our method for creating scale-specific neurons in deep NLP models constrains how the activation of a neuron can change across the tokens of an input by interpreting those activations as a digital signal and filtering out parts of its frequency spectrum. This technique enables us to extract scale-specific information from BERT representations: by filtering out different frequencies we can produce new representations that perform well on part of speech tagging (word-level), dialog speech acts classification (utterance-level), or topic classification (document-level), while performing poorly on the other tasks. We also present a prism layer for use during training, which constrains different neurons of a BERT model to different parts of the frequency spectrum. Our proposed BERT + Prism model is better able to predict masked tokens using long-range context, and produces individual multiscale representations that perform with comparable or improved performance across all three tasks. Our methods are general and readily applicable to other domains besides language, such as images, audio, and video.
**********

DVERGE: Diversifying Vulnerabilities for Enhanced Robust Generation of Ensembles
Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, Hai Li
Recent research finds CNN models for image classification demonstrate overlapped adversarial vulnerabilities: adversarial attacks can mislead CNN models with small perturbations, which can effectively transfer between different models trained on the same dataset. Adversarial training, as a general robustness improvement technique, eliminates the vulnerability in a single model by forcing it to learn robust features. The process is hard, often requires models with large capacity, and suffers from significant loss on clean data accuracy. Alternatively, ensemble methods are proposed to induce sub-models with diverse outputs against a transfer adversarial example, making the ensemble robust against transfer attacks even if each sub-model is individually non-robust. Only small clean accuracy drop is observed in the process. However, previous ensemble training methods are not efficacious in inducing such diversity and thus ineffective on reaching robust ensemble. We propose DVERGE, which isolates the adversarial vulnerability in each sub-model by distilling non-robust features, and diversifies the adversarial vulnerability to induce diverse outputs against a transfer attack. The novel diversity metric and training procedure enables DVERGE to achieve higher robustness against transfer attacks comparing to previous ensemble methods, and enables the improved robustness when more sub-models are added to the ensemble. The code of this work is available at https://github.com/zjysteven/DVERGE.
**********

Towards practical differentially private causal graph discovery
Lun Wang, Qi Pang, Dawn Song
Causal graph discovery refers to the process of discovering causal relation graphs
from purely observational data. Like other statistical data, a causal graph might
leak sensitive information about participants in the dataset. In this paper, we present a differentially private causal graph discovery algorithm, Priv-PC, which
improves both utility and running time compared to the state-of-the-art. The design of Priv-PC follows a novel paradigm called sieve-and-examine which
uses a small amount of privacy budget to filter out “insignificant” queries, and leverages the remaining budget to obtain highly accurate answers for the “significant” queries. We also conducted the first sensitivity analysis for conditional independence tests including conditional Kendall’s $\tau$ and conditional Spearman’s $\rho$. We evaluated Priv-PC on 7 public datasets and compared with the
state-of-the-art. The results show that Priv-PC achieves 10.61 to 293.87 times speedup and better utility. The implementation of Priv-PC, including the code used in our evaluation, is available at https://github.com/sunblaze-ucb/

Priv-PC-Differentially-Private-Causal-Graph-Discovery.
**********
Independent Policy Gradient Methods for Competitive Reinforcement Learning
Constantinos Daskalakis, Dylan J. Foster, Noah Golowich
We obtain global, non-asymptotic convergence guarantees for independent learning algorithms in competitive reinforcement learning settings with two agents (i.e., zero-sum stochastic games). We consider an episodic setting where in each episode, each player independently selects a policy and observes only their own actions and rewards, along with the state. We show that if both players run policy gradient methods in tandem, their policies will converge to a min-max equilibrium of the game, as long as their learning rates follow a two-timescale rule (which is necessary). To the best of our knowledge, this constitutes the first finite-sample convergence result for independent learning in competitive RL, as prior work has largely focused on centralized/coordinated procedures for equilibrium computation.
**********
The Value Equivalence Principle for Model-Based Reinforcement Learning
Christopher Grimm, Andre Barreto, Satinder Singh, David Silver
Learning models of the environment from data is often viewed as an essential component to building intelligent reinforcement learning (RL) agents. The common practice is to separate the learning of the model from its use, by constructing a model of the environment’s dynamics that correctly predicts the observed state transitions.  In this paper we argue that the limited representational resources of model-based RL agents are better used to build models that are directly useful for value-based planning.  As our main contribution, we introduce the principle of value equivalence: two models are value equivalent with respect to a set of functions and policies if they yield the same Bellman updates. We propose a formulation of the model learning problem based on the value equivalence principle and analyze how the set of feasible solutions is impacted by the choice of policies and functions. Specifically, we show that, as we augment the set of policies and functions considered, the class of value equivalent models shrinks, until eventually collapsing to a single point corresponding to a model that perfectly describes the environment. In many problems, directly modelling state-to-state transitions may be both difficult and unnecessary. By leveraging the value-equivalence principle one may find simpler models without compromising performance, saving computation and memory.  We illustrate the benefits of value-equivalent model learning with experiments comparing it against more traditional counterparts like maximum likelihood estimation.  More generally, we argue that the principle of value equivalence underlies a number of recent empirical successes in RL, such as Value Iteration Networks, the Predictron, Value Prediction Networks, TreeQN, and MuZero, and provides a first theoretical underpinning of those results.
**********
Structured Convolutions for Efficient Neural Network Design
Yash Bhalgat, Yizhe Zhang, Jamie Menjay Lin, Fatih Porikli
In this work, we tackle model efficiency by exploiting redundancy in the implicit structure of the building blocks of convolutional neural networks. We start our analysis by introducing a general definition of Composite Kernel structures that enable the execution of convolution operations in the form of efficient, scaled, sum-pooling components. As its special case, we propose Structured Convolutions and show that these allow decomposition of the convolution operation into a sum-pooling operation followed by a convolution with significantly lower complexity and fewer weights. We show how this decomposition can be applied to 2D and 3D kernels as well as the fully-connected layers. Furthermore, we present a Structural Regularization loss that promotes neural network layers to leverage on this desired structure in a way that, after training, they can be decomposed with negligible performance loss. By applying our method to a wide range of CNN architectures, we demonstrate 'structured' versions of the ResNets that are up to 2x smaller and a new Structured-MobileNetV2 that is more efficient while staying within an accuracy loss of 1% on ImageNet and CIFAR-10 datasets. We also show similar structured versions of EfficientNet on ImageNet and HRNet architecture for se

mantic segmentation on the Cityscapes dataset. Our method performs equally well or superior in terms of the complexity reduction in comparison to the existing tensor decomposition and channel pruning methods.
**********

Latent World Models For Intrinsically Motivated Exploration
Aleksandr Ermolov, Nicu Sebe
In this work we consider partially observable environments with sparse rewards. We present a self-supervised representation learning method for image-based observations, which arranges embeddings respecting temporal distance of observations. This representation is empirically robust to stochasticity and suitable for novelty detection from the error of a predictive forward model. We consider episodic and life-long uncertainties to guide the exploration. We propose to estimate the missing information about the environment with the world model, which operates in the learned latent space. As a motivation of the method, we analyse the exploration problem in a tabular Partially Observable Labyrinth. We demonstrate the method on image-based hard exploration environments from the Atari benchmark and report significant improvement with respect to prior work. The source code of the method and all the experiments is available at https://github.com/htdt/lwm.
**********

Estimating Rank-One Spikes from Heavy-Tailed Noise via Self-Avoiding Walks
Jingqiu Ding, Samuel Hopkins, David Steurer
We study symmetric spiked matrix models with respect to a general class of noise distributions. Given a rank-1 deformation of a random noise matrix, whose entries
are independently distributed with zero mean and unit variance, the goal is to estimate the rank-1 part. For the case of Gaussian noise, the top eigenvector of the given matrix is a widely-studied estimator known to achieve optimal statistical
guarantees, e.g., in the sense of the celebrated BBP phase transition. However, this
estimator can fail completely for heavy-tailed noise.
**********

Policy Improvement via Imitation of Multiple Oracles
Ching-An Cheng, Andrey Kolobov, Alekh Agarwal
Despite its promise, reinforcement learning’s real-world adoption has been hampered by the need for costly exploration to learn a good policy. Imitation learning (IL) mitigates this shortcoming by using an oracle policy during training as a bootstrap to accelerate the learning process. However, in many practical situations, the learner has access to multiple suboptimal oracles, which may provide con■icting advice in a state. The existing IL literature provides a limited treatment of such scenarios. Whereas in the single-oracle case, the return of the oracle’s policy provides an obvious benchmark for the learner to compete against, neither such a benchmark nor principled ways of outperforming it are known for the multi-oracle setting. In this paper, we propose the state-wise maximum of the oracle policies’ values as a natural baseline to resolve con■icting advice from multiple oracles. Using a reduction of policy optimization to online learning, we introduce a novel IL algorithm MAMBA, which can provably learn a policy competitive with this benchmark. In particular, MAMBA optimizes policies by using a gradient estimator in the style of generalized advantage estimation (GAE). Our theoretical analysis shows that this design makes MAMBA robust and enables it to outperform the oracle policies by a larger margin than the IL state of the art, even in the single-oracle case. In an evaluation against standard policy gradient with GAE and AggreVaTe(D), we showcase MAMBA’s ability to leverage demonstrations both from a single and from multiple weak oracles, and signi■cantly speed up policy optimization.
**********

Training Generative Adversarial Networks by Solving Ordinary Differential Equations
Chongli Qin, Yan Wu, Jost Tobias Springenberg, Andy Brock, Jeff Donahue, Timothy Lillicrap, Pushmeet Kohli

The instability of Generative Adversarial Network (GAN) training has frequently been attributed to gradient descent. Consequently, recent methods have aimed to tailor the models and training procedures to stabilise the discrete updates. In contrast, we study the continuous-time dynamics induced by GAN training. Both th eory and toy experiments suggest that these dynamics are in fact surprisingly st able. From this perspective, we hypothesise that instabilities in training GANs arise from the integration error in discretising the continuous dynamics. We exp erimentally verify that well-known ODE solvers (such as Runge-Kutta) can stabili se training - when combined with a regulariser that controls the integration err or. Our approach represents a radical departure from previous methods which typi cally use adaptive optimisation and stabilisation techniques that constrain the functional space (e.g. Spectral Normalisation). Evaluation on CIFAR-10 and Image Net shows that our method outperforms several strong baselines, demonstrating it s efficacy.

**********

Learning of Discrete Graphical Models with Neural Networks

Abhijith Jayakumar, Andrey Lokhov, Sidhant Misra, Marc Vuffray

Graphical models are widely used in science to represent joint probability distr ibutions with an underlying conditional dependence structure. The inverse proble m of learning a discrete graphical model given i.i.d samples from its joint dist ribution can be solved with near-optimal sample complexity using a convex optimi zation method known as Generalized Regularized Interaction Screening Estimator ( GRISE). But the computational cost of GRISE becomes prohibitive when the energy function of the true graphical model has higher order terms. We introduce NeurIS E, a neural net based algorithm for graphical model learning, to tackle this lim itation of GRISE. We use neural nets as function approximators in an Interaction Screening objective function. The optimization of this objective then produces a neural-net representation for the conditionals of the graphical model. NeurIS E algorithm is seen to be a better alternative to GRISE when the energy function of the true model has a high order with a high degree of symmetry. In these cas es NeurISE is able to find the correct parsimonious representation for the condi tionals without being fed any prior information about the true model. NeurISE ca n also be used to learn the underlying structure of the true model with some sim ple modifications to its training procedure. In addition, we also show a variant of NeurISE that can be used to learn a neural net representation for the full energy function of the true model.

**********

RepPoints v2: Verification Meets Regression for Object Detection

Yihong Chen, Zheng Zhang, Yue Cao, Liwei Wang, Stephen Lin, Han Hu

Verification and regression are two general methodologies for prediction in neur al networks. Each has its own strengths: verification can be easier to infer acc urately, and regression is more efficient and applicable to continuous target va riables. Hence, it is often beneficial to carefully combine them to take advanta ge of their benefits. In this paper, we take this philosophy to improve state-of -the-art object detection, specifically by RepPoints. Though RepPoints provides high performance, we find that its heavy reliance on regression for object local ization leaves room for improvement. We introduce verification tasks into the lo calization prediction of RepPoints, producing RepPoints v2, which proves consist ent improvements of about 2.0 mAP over the original RepPoints on COCO object det ection benchmark using different backbones and training methods. RepPoints v2 al so achieves 52.1 mAP on the COCO \texttt{test-dev} by a single model. Moreover, we show that the proposed approach can more generally elevate other object detec tion frameworks as well as applications such as instance segmentation.

**********

Unfolding the Alternating Optimization for Blind Super Resolution

zhengxiong luo, Yan Huang, Shang Li, Liang Wang, Tieniu Tan

Previous methods decompose blind super resolution (SR) problem into two sequenti al steps: \textit{i}) estimating blur kernel from given low-resolution (LR) imag e and \textit{ii}) restoring SR image based on estimated kernel. This two-step s olution involves two independently trained models, which may not well compatible

with each other. Small estimation error of the first step could cause severe pe rformance drop of the second one. While on the other hand, the first step can on ly utilize limited information from LR image, which makes it difficult to predic t highly accurate blur kernel. Towards these issues, instead of considering thes e two steps separately, we adopt an alternating optimization algorithm, which ca n estimate blur kernel and restore SR image in a single model. Specifically, we design two convolutional neural modules, namely \textit{Restorer} and \textit{E stimator}. \textit{Restorer} restores SR image based on predicted kernel, and \t extit{Estimator} estimates blur kernel with the help of restored SR image. We a lternate these two modules repeatedly and unfold this process to form an end-to-end trainable network. In this way, \textit{Estimator} utilizes information from both LR and SR images, which makes the estimation of blur kernel easier. More i mportantly, \textit{Restorer} is trained with the kernel estimated by \textit{Es timator}, instead of ground-truth kernel, thus \textit{Restorer} could be more t olerant to the estimation error of \textit{Estimator}. Extensive experiments on synthetic datasets and real-world images show that our model can largely outperf orm state-of-the-art methods and produce more visually favorable results at much higher speed. The source code will be publicly available.

**********

Entrywise convergence of iterative methods for eigenproblems

Vasileios Charisopoulos, Austin R. Benson, Anil Damle

Several problems in machine learning, statistics, and other fields rely on compu ting eigenvectors. For large scale problems, the computation of these eigenvecto rs is typically performed via iterative schemes such as subspace iteration or Kr ylov methods. While there is classical and comprehensive analysis for subspace c onvergence guarantees with respect to the spectral norm, in many modern applicat ions other notions of subspace distance are more appropriate. Recent theoretical work has focused on perturbations of subspaces measured in the ■2→∞ norm, but d oes not consider the actual computation of eigenvectors. Here we address the con vergence of subspace iteration when distances are measured in the ■2→∞ norm and provide deterministic bounds. We complement our analysis with a practical stoppi ng criterion and demonstrate its applicability via numerical experiments. Our re sults show that one can get comparable performance on downstream tasks while req uiring fewer iterations, thereby saving substantial computational time.

**********

Learning Object-Centric Representations of Multi-Object Scenes from Multiple Vie ws

Nanbo Li, Cian Eastwood, Robert Fisher

Learning object-centric representations of multi-object scenes is a promising ap proach towards machine intelligence, facilitating high-level reasoning and contr ol from visual sensory data. However, current approaches for \textit{unsupervise d object-centric scene representation} are incapable of aggregating information from multiple observations of a scene. As a result, these ``single-view'' method s form their representations of a 3D scene based only on a single 2D observation (view). Naturally, this leads to several inaccuracies, with these methods falli ng victim to single-view spatial ambiguities. To address this, we propose \texti t{The Multi-View and Multi-Object Network (MulMON)}---a method for learning accu rate, object-centric representations of multi-object scenes by leveraging multip le views. In order to sidestep the main technical difficulty of the \textit{mult i-object-multi-view} scenario---maintaining object correspondences across views---MulMON iteratively updates the latent object representations for a scene over multiple views. To ensure that these iterative updates do indeed aggregate spati al information to form a complete 3D scene understanding, MulMON is asked to pre dict the appearance of the scene from novel viewpoints during training. Through experiments we show that MulMON better-resolves spatial ambiguities than single-view methods---learning more accurate and disentangled object representations---and also achieves new functionality in predicting object segmentations for novel viewpoints.

**********

A Catalyst Framework for Minimax Optimization

Junchi Yang, Siqi Zhang, Negar Kiyavash, Niao He
We introduce a generic \emph{two-loop} scheme for smooth minimax optimization wi th strongly-convex-concave objectives. Our approach applies the accelerated prox imal point framework (or Catalyst) to the associated \emph{dual problem} and tak es full advantage of existing gradient-based algorithms to solve a sequence of w ell-balanced strongly-convex-strongly-concave minimax problems. Despite its simp licity, this leads to a family of near-optimal algorithms with improved complexi ty over all existing methods designed for strongly-convex-concave minimax proble ms. Additionally, we obtain the first variance-reduced algorithms for this class of minimax problems with finite-sum structure and establish even faster converg ence rate. Furthermore, when extended to the nonconvex-concave minimax optimizat ion, our algorithm again achieves the state-of-the-art complexity for finding a stationary point. We carry out several numerical experiments showcasing the supe riority of the Catalyst framework in practice.
**********

Self-supervised Co-Training for Video Representation Learning
Tengda Han, Weidi Xie, Andrew Zisserman
The objective of this paper is visual-only self-supervised video representation learning. We make the following contributions: (i) we investigate the benefit of adding semantic-class positives to instance-based Info Noise Contrastive Estima tion (InfoNCE) training, showing that this form of supervised contrastive learni ng leads to a clear improvement in performance; (ii) we propose a novel self-sup ervised co-training scheme to improve the popular infoNCE loss, exploiting the c omplementary information from different views, RGB streams and optical flow, of the same data source by using one view to obtain positive class samples for the other; (iii) we thoroughly evaluate the quality of the learnt representation on two different downstream tasks: action recognition and video retrieval. In both cases, the proposed approach demonstrates state-of-the-art or comparable perform ance with other self-supervised approaches, whilst being significantly more effi cient to train, i.e. requiring far less training data to achieve similar perform ance.
**********

Gradient Estimation with Stochastic Softmax Tricks
Max Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, Chris J. Maddison
The Gumbel-Max trick is the basis of many relaxed gradient estimators. These est imators are easy to implement and low variance, but the goal of scaling them com prehensively to large combinatorial distributions is still outstanding. Working within the perturbation model framework, we introduce stochastic softmax tricks, which generalize the Gumbel-Softmax trick to combinatorial spaces. Our framewor k is a unified perspective on existing relaxed estimators for perturbation model s, and it contains many novel relaxations. We design structured relaxations for subset selection, spanning trees, arborescences, and others. When compared to le ss structured baselines, we find that stochastic softmax tricks can be used to t rain latent variable models that perform better and discover more latent structu re.
**********

Meta-Learning Requires Meta-Augmentation
Janarthanan Rajendran, Alexander Irpan, Eric Jang
Meta-learning algorithms aim to learn two components: a model that predicts targ ets for a task, and a base learner that updates that model when given examples f rom a new task. This additional level of learning can be powerful, but it also c reates another potential source of overfitting, since we can now overfit in eith er the model or the base learner. We describe both of these forms of meta-learni ng overfitting, and demonstrate that they appear experimentally in common meta-l earning benchmarks. We introduce an information-theoretic framework of meta-augm entation, whereby adding randomness discourages the base learner and model from learning trivial solutions that do not generalize to new tasks. We demonstrate t hat meta-augmentation produces large complementary benefits to recently proposed meta-regularization techniques.
**********

SLIP: Learning to predict in unknown dynamical systems with long-term memory
Paria Rashidinejad, Jiantao Jiao, Stuart Russell
We present an efficient and practical (polynomial time) algorithm for online prediction in unknown and partially observed linear dynamical systems (LDS) under stochastic noise. When the system parameters are known, the optimal linear predictor is the Kalman filter. However, in unknown systems, the performance of existing predictive models is poor in important classes of LDS that are only marginally stable and exhibit long-term forecast memory. We tackle this problem by bounding the generalized Kolmogorov width of the Kalman filter coefficient set. This motivates the design of an algorithm, which we call spectral LDS improper predictor (SLIP), based on conducting a tight convex relaxation of the Kalman predictive model via spectral methods. We provide a finite-sample analysis, showing that our algorithm competes with the Kalman filter in hindsight with only logarithmic regret. Our regret analysis relies on Mendelson's small-ball method, providing sharp error bounds without concentration, boundedness, or exponential forgetting assumptions. Empirical evaluations demonstrate that SLIP outperforms state-of-the-art methods in LDS prediction. Our theoretical and experimental results shed light on the conditions required for efficient probably approximately correct (PAC) learning of the Kalman filter from partially observed data.
**********
Improving GAN Training with Probability Ratio Clipping and Sample Reweighting
Yue Wu, Pan Zhou, Andrew G. Wilson, Eric Xing, Zhiting Hu
Despite success on a wide range of problems related to vision, generative adversarial networks (GANs) often suffer from inferior performance due to unstable training, especially for text generation. To solve this issue, we propose a new variational GAN training framework which enjoys superior training stability. Our approach is inspired by a connection of GANs and reinforcement learning under a variational perspective. The connection leads to (1) probability ratio clipping that regularizes generator training to prevent excessively large updates, and (2) a sample re-weighting mechanism that improves discriminator training by downplaying bad-quality fake samples. Moreover, our variational GAN framework can provably overcome the training issue in many GANs that an optimal discriminator cannot provide any informative gradient to training generator. By plugging the training approach in diverse state-of-the-art GAN architectures, we obtain significantly improved performance over a range of tasks, including text generation, text style transfer, and image generation.
**********
Bayesian Bits: Unifying Quantization and Pruning
Mart van Baalen, Christos Louizos, Markus Nagel, Rana Ali Amjad, Ying Wang, Tijmen Blankevoort, Max Welling
We introduce Bayesian Bits, a practical method for joint mixed precision quantization and pruning through gradient based optimization. Bayesian Bits employs a novel decomposition of the quantization operation, which sequentially considers doubling the bit width. At each new bit width, the residual error between the full precision value and the previously rounded value is quantized. We then decide whether or not to add this quantized residual error for a higher effective bit width and lower quantization noise. By starting with a power-of-two bit width, this decomposition will always produce hardware-friendly configurations, and through an additional 0-bit option, serves as a unified view of pruning and quantization. Bayesian Bits then introduces learnable stochastic gates, which collectively control the bit width of the given tensor. As a result, we can obtain low bit solutions by performing approximate inference over the gates, with prior distributions that encourage most of them to be switched off. We experimentally validate our proposed method on several benchmark datasets and show that we can learn pruned, mixed precision networks that provide a better trade-off between accuracy and efficiency than their static bit width equivalents.
**********
On Testing of Samplers
Kuldeep S Meel, Yash Pralhad Pote, Sourav Chakraborty
Given a set of items F and a weight function W: F -> (0,1), the problem of sampl

ing seeks to sample an item proportional to its weight. Sampling is a fundamental problem in machine learning. The daunting computational complexity of sampling with formal guarantees leads designers to propose heuristics-based techniques for which no rigorous theoretical analysis exists to quantify the quality of the generated distributions.

This poses a challenge in designing a testing methodology to test whether a sampler under test generates samples according to a given distribution. Only recently, Chakraborty and Meel (2019) designed the first scalable verifier, called Barbarik, for samplers in the special case when the weight function W is constant, that is, when the sampler is supposed to sample uniformly from F. The techniques in Barbarik, however, fail to handle general weight functions.

The primary contribution of this paper is an affirmative answer to the above challenge: motivated by Barbarik, but using different techniques and analysis, we design Barbarik2, an algorithm to test whether the distribution generated by a sampler is epsilon-close or eta-far from any target distribution. In contrast to black-box sampling techniques that require a number of samples proportional to |F|, Barbarik2 requires only \tilde{O}(Tilt(W, F)^2/eta(eta - 6*epsilon)^3) samples, where the Tilt is the maximum ratio of weights of two points in F. Barbarik2 can handle any arbitrary weight function. We present a prototype implementation of Barbarik2 and use it to test three state-of-the-art samplers.

**********

Gaussian Process Bandit Optimization of the Thermodynamic Variational Objective

Vu Nguyen, Vaden Masrani, Rob Brekelmans, Michael Osborne, Frank Wood

Achieving the full promise of the Thermodynamic Variational Objective (TVO), a recently proposed variational inference objective that lower-bounds the log evidence via one-dimensional Riemann integration, requires choosing a ``schedule'' of sorted discretization points. This paper introduces a bespoke Gaussian process bandit optimization method for automatically choosing these points. Our approach not only automates their one-time selection, but also dynamically adapts their positions over the course of optimization, leading to improved model learning and inference. We provide theoretical guarantees that our bandit optimization converges to the regret-minimizing choice of integration points. Empirical validation of our algorithm is provided in terms of improved learning and inference in Variational Autoencoders and sigmoid belief networks.

**********

MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, Ming Zhou

Pre-trained language models (e.g., BERT (Devlin et al., 2018) and its variants) have achieved remarkable success in varieties of NLP tasks. However, these models usually consist of hundreds of millions of parameters which brings challenges for fine-tuning and online serving in real-life applications due to latency and capacity constraints. In this work, we present a simple and effective approach to compress large Transformer (Vaswani et al., 2017) based pre-trained models, termed as deep self-attention distillation. The small model (student) is trained by deeply mimicking the self-attention module, which plays a vital role in Transformer networks, of the large model (teacher). Specifically, we propose distilling the self-attention module of the last Transformer layer of the teacher, which is effective and flexible for the student. Furthermore, we introduce the scaled dot-product between values in the self-attention module as the new deep self-attention knowledge, in addition to the attention distributions (i.e., the scaled dot-product of queries and keys) that have been used in existing works. Moreover, we show that introducing a teacher assistant (Mirzadeh et al., 2019) also helps the distillation of large pre-trained Transformer models. Experimental results demonstrate that our monolingual model outperforms state-of-the-art baselines in different parameter size of student models. In particular, it retains more than 99% accuracy on SQuAD 2.0 and several GLUE benchmark tasks using 50% of the Transformer parameters and computations of the teacher model. We also obtain competitive results in applying deep self-attention distillation to multilingual pre-trained models.

**********
Optimal Epoch Stochastic Gradient Descent Ascent Methods for Min-Max Optimizatio
n
Yan Yan, Yi Xu, Qihang Lin, Wei Liu, Tianbao Yang
Epoch gradient descent method (a.k.a. Epoch-GD) proposed by (Hazan and Kale, 201
1) was deemeda breakthrough for stochastic strongly convex minimization, which a
chieves theoptimal convergence rate of O(1/T) with T iterative updates for the o
bjective gap. However, its extension to solving stochastic min-max problems with
 strong convexity and strong concavity still remains open, and it is still uncle
ar whethera fast rate ofO(1/T)for theduality gapis achievable for stochastic min
-max optimization under strong convexity and strong concavity.  Although some re
-cent studies have proposed stochastic algorithms with fast convergence rates fo
rmin-max problems, they require additional assumptions about the problem, e.g.,s
moothness, bi-linear structure, etc. In this paper, we bridge this gap by provid
inga sharp analysis of epoch-wise stochastic gradient descent ascent method (ref
erredto as Epoch-GDA) for solving strongly convex strongly concave (SCSC) min-ma
xproblems, without imposing any additional assumption about smoothness or the fu
nction’s structure. To the best of our knowledge, our result is the first one th
at shows Epoch-GDA can achieve the optimal rate ofO(1/T)for the duality gapof ge
neral SCSC min-max problems. We emphasize that such generalization of Epoch-GD f
or strongly convex minimization problems to Epoch-GDA for SCSC min-max problems
is non-trivial and requires novel technical analysis. Moreover, we notice that t
he key lemma can also be used for proving the convergence of Epoch-GDA for weakl
y-convex strongly-concave min-max problems, leading to a nearly optimal complexi
ty without resorting to smoothness or other structural conditions.
**********
Woodbury Transformations for Deep Generative Flows
You Lu, Bert Huang
Normalizing flows are deep generative models that allow efficient likelihood cal
culation and sampling. The core requirement for this advantage is that they are
constructed using functions that can be efficiently inverted and for which the d
eterminant of the function's Jacobian can be efficiently computed. Researchers h
ave introduced various such flow operations, but few of these allow rich interac
tions among variables without incurring significant computational costs. In this
 paper, we introduce Woodbury transformations, which achieve efficient invertibi
lity via the Woodbury matrix identity and efficient determinant calculation via
Sylvester's determinant identity. In contrast with other operations used in stat
e-of-the-art normalizing flows, Woodbury transformations enable (1) high-dimensi
onal interactions, (2) efficient sampling, and (3) efficient likelihood evaluati
on. Other similar operations, such as 1x1 convolutions, emerging convolutions, o
r periodic convolutions allow at most two of these three advantages. In our expe
riments on multiple image datasets, we find that Woodbury transformations allow
learning of higher-likelihood models than other flow architectures while still e
njoying their efficiency advantages.
**********
Graph Contrastive Learning with Augmentations
Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, Yang Shen
Generalizable, transferrable, and robust representation learning on graph-struct
ured data remains a challenge for current graph neural networks (GNNs). Unlike w
hat has been developed for convolutional neural networks (CNNs) for image data,
self-supervised learning and pre-training are less explored for GNNs. In this pa
per, we propose a graph contrastive learning (GraphCL) framework for learning un
supervised representations of graph data. We first design four types of graph au
gmentations to incorporate various priors. We then systematically study the impa
ct of various combinations of graph augmentations on multiple datasets, in four
different settings: semi-supervised, unsupervised, and transfer learning as well
 as adversarial attacks. The results show that, even without tuning augmentation
 extents nor using sophisticated GNN architectures, our GraphCL framework can pr
oduce graph representations of similar or better generalizability, transferrabil
ity, and robustness compared to state-of-the-art methods. We also investigate th

e impact of parameterized graph augmentation extents and patterns, and observe f urther performance gains in preliminary experiments. Our codes are available at https://github.com/Shen-Lab/GraphCL.
**********
Gradient Surgery for Multi-Task Learning
Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, Chelsea Finn
While deep learning and deep reinforcement learning (RL) systems have demonstrat ed impressive results in domains such as image classification, game playing, and robotic control, data efficiency remains a major challenge. Multi-task learning has emerged as a promising approach for sharing structure across multiple tasks to enable more efficient learning. However, the multi-task setting presents a n umber of optimization challenges, making it difficult to realize large efficienc y gains compared to learning tasks independently. The reasons why multi-task lea rning is so challenging compared to single-task learning are not fully understoo d. In this work, we identify a set of three conditions of the multi-task optimiz ation landscape that cause detrimental gradient interference, and develop a simp le yet general approach for avoiding such interference between task gradients. W e propose a form of gradient surgery that projects a task's gradient onto the no rmal plane of the gradient of any other task that has a gradient. On a series of challenging multi-task supervised and multi-task RL problems, this approach lea ds to substantial gains in efficiency and performance. Further, it is model-agno stic and can be combined with previously-proposed multi-task architectures for e nhanced performance.
**********
Bayesian Probabilistic Numerical Integration with Tree-Based Models
Harrison Zhu, Xing Liu, Ruya Kang, Zhichao Shen, Seth Flaxman, Francois-Xavier B riol
Bayesian quadrature (BQ) is a method for solving numerical integration problems in a Bayesian manner, which allows users to quantify their uncertainty about the solution. The standard approach to BQ is based on a Gaussian process (GP) appro ximation of the integrand. As a result, BQ is inherently limited to cases where GP approximations can be done in an efficient manner, thus often prohibiting ver y high-dimensional or non-smooth target functions. This paper proposes to tackle this issue with a new Bayesian numerical integration algorithm based on Bayesia n Additive Regression Trees (BART) priors, which we call BART-Int. BART priors a re easy to tune and well-suited for discontinuous functions. We demonstrate that they also lend themselves naturally to a sequential design setting and that exp licit convergence rates can be obtained in a variety of settings. The advantages and disadvantages of this new methodology are highlighted on a set of benchmark tests including the Genz functions, on a rare-event simulation problem and on a Bayesian survey design problem.
**********
Deep learning versus kernel learning: an empirical study of loss landscape geome try and the time evolution of the Neural Tangent Kernel
Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, D aniel M. Roy, Surya Ganguli
In suitably initialized wide networks, small learning rates transform deep neura l networks (DNNs) into neural tangent kernel (NTK) machines, whose training dyna mics is well-approximated by a linear weight expansion of the network at initial ization.  Standard training, however, diverges from its linearization in ways th at are poorly understood. We study the relationship between the training dynamic s of nonlinear deep networks, the geometry of the loss landscape, and the time e volution of a data-dependent NTK. We do so through a large-scale phenomenologica l analysis of training, synthesizing diverse measures characterizing loss landsc ape geometry and NTK dynamics. In multiple neural architectures and datasets, we find these diverse measures evolve in a highly correlated manner, revealing a u niversal picture of the deep learning process.  In this picture, deep network tr aining exhibits a highly chaotic rapid initial transient that within 2 to 3 epoc hs determines the final linearly connected basin of low loss containing the end

point of training. During this chaotic transient, the NTK changes rapidly, learning useful features from the training data that enables it to outperform the standard initial NTK by a factor of 3 in less than 3 to 4 epochs. After this rapid chaotic transient, the NTK changes at constant velocity, and its performance matches that of full network training in 15\% to 45\% of training time. Overall, our analysis reveals a striking correlation between a diverse set of metrics over training time, governed by a rapid chaotic to stable transition in the first few epochs, that together poses challenges and opportunities for the development of more accurate theories of deep learning.

**********

Graph Meta Learning via Local Subgraphs

Kexin Huang, Marinka Zitnik

Prevailing methods for graphs require abundant label and edge information for learning. When data for a new task are scarce, meta-learning can learn from prior experiences and form much-needed inductive biases for fast adaption to new tasks. Here, we introduce G-Meta, a novel meta-learning algorithm for graphs. G-Meta uses local subgraphs to transfer subgraph-specific information and learn transferable knowledge faster via meta gradients. G-Meta learns how to quickly adapt to a new task using only a handful of nodes or edges in the new task and does so by learning from data points in other graphs or related, albeit disjoint label sets. G-Meta is theoretically justified as we show that the evidence for a prediction can be found in the local subgraph surrounding the target node or edge. Experiments on seven datasets and nine baseline methods show that G-Meta outperforms existing methods by up to 16.3%. Unlike previous methods, G-Meta successfully learns in challenging, few-shot learning settings that require generalization to completely new graphs and never-before-seen labels. Finally, G-Meta scales to large graphs, which we demonstrate on a new Tree-of-Life dataset comprising of 1,840 graphs, a two-orders of magnitude increase in the number of graphs used in prior work.

**********

Stochastic Deep Gaussian Processes over Graphs

Naiqi Li, Wenjie Li, Jifeng Sun, Yinghua Gao, Yong Jiang, Shu-Tao Xia

In this paper we propose Stochastic Deep Gaussian Processes over Graphs (DGPG), which are deep structure models that learn the mappings between input and output signals in graph domains. The approximate posterior distributions of the latent variables are derived with variational inference, and the evidence lower bound is evaluated and optimized by the proposed recursive sampling scheme. The Bayesian non-parametric natural of our model allows it to resist overfitting, while the expressive deep structure grants it the potential to learn complex relations. Extensive experiments demonstrate that our method achieves superior performances in both small size (< 50) and large size (> 35,000) datasets. We show that DGPG outperforms another Gaussian-based approach, and is competitive to a state-of-the-art method in the challenging task of traffic flow prediction. Our model is also capable of capturing uncertainties in a mathematical principled way and automatically discovering which vertices and features are relevant to the prediction.

**********

Bayesian Causal Structural Learning with Zero-Inflated Poisson Bayesian Networks

Junsouk Choi, Robert Chapkin, Yang Ni

Multivariate zero-inflated count data arise in a wide range of areas such as economics, social sciences, and biology. To infer causal relationships in zero-inflated count data, we propose a new zero-inflated Poisson Bayesian network (ZIPBN) model. We show that the proposed ZIPBN is identifiable with cross-sectional data. The proof is based on the well-known characterization of Markov equivalence class which is applicable to other distribution families. For causal structural learning, we introduce a fully Bayesian inference approach which exploits the parallel tempering Markov chain Monte Carlo algorithm to efficiently explore the multi-modal network space. We demonstrate the utility of the proposed ZIPBN in causal discoveries for zero-inflated count data by simulation studies with comparison to alternative Bayesian network methods. Additionally, real sin

gle-cell RNA-sequencing data with known causal relationships will be used to ass
ess the capability of ZIPBN for discovering causal relationships in real-world p
roblems.
**********

Evaluating Attribution for Graph Neural Networks
Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Wang, Wes
ley Qian, Kevin McCloskey, Lucy Colwell , Alexander Wiltschko
Interpretability of machine learning models is critical to scientific understand
ing, AI safety, as well as debugging. Attribution is one approach to interpretab
ility, which highlights input dimensions that are influential to a neural networ
k’s prediction. Evaluation of these methods is largely qualitative for image and
 text models, because acquiring ground truth attributions requires expensive and
 unreliable human judgment. Attribution has been little studied for graph neural
 networks (GNNs), a model class of growing importance that makes predictions on
arbitrarily-sized graphs. In this work we adapt commonly-used attribution method
s for GNNs and quantitatively evaluate them using computable ground-truths that
are objective and challenging to learn. We make concrete recommendations for whi
ch attribution methods to use, and provide the data and code for our benchmarkin
g suite. Rigorous and open source benchmarking of attribution methods in graphs
could enable new methods development and broader use of attribution in real-worl
d ML tasks.
**********

On Second Order Behaviour in Augmented Neural ODEs
Alexander Norcliffe, Cristian Bodnar, Ben Day, Nikola Simidjievski, Pietro Lió
Neural Ordinary Differential Equations (NODEs) are a new class of models that
transform data continuously through infinite-depth architectures. The continuous
nature of NODEs has made them particularly suitable for learning the dynamics
of complex physical systems. While previous work has mostly been focused on
first order ODEs, the dynamics of many systems, especially in classical physics,
are governed by second order laws. In this work, we consider Second Order
Neural ODEs (SONODEs). We show how the adjoint sensitivity method can be
extended to SONODEs and prove that the optimisation of a first order coupled
ODE is equivalent and computationally more efficient. Furthermore, we extend the
theoretical understanding of the broader class of Augmented NODEs (ANODEs)
by showing they can also learn higher order dynamics with a minimal number
of augmented dimensions, but at the cost of interpretability. This indicates tha
t
the advantages of ANODEs go beyond the extra space offered by the augmented
dimensions, as originally thought. Finally, we compare SONODEs and ANODEs
on synthetic and real dynamical systems and demonstrate that the inductive biase
s
of the former generally result in faster training and better performance.
**********

Neuron Shapley: Discovering the Responsible Neurons
Amirata Ghorbani, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Stochastic Normalizing Flows
Hao Wu, Jonas Köhler, Frank Noe
The sampling of probability distributions specified up to a normalization consta
nt is an important problem in both machine learning and statistical mechanics. W
hile classical stochastic sampling methods such as Markov Chain Monte Carlo (MCM
C) or Langevin Dynamics (LD) can suffer from slow mixing times there is a growin
g interest in using normalizing flows in order to learn the transformation of a
simple prior distribution to the given target distribution. Here we propose a ge
neralized and combined approach to sample target densities: Stochastic Normalizi
ng Flows (SNF) – an arbitrary sequence of deterministic invertible functions and

stochastic sampling blocks. We show that stochasticity overcomes expressivity limitations of normalizing flows resulting from the invertibility constraint, whereas trainable transformations between sampling steps improve efficiency of pure MCMC/LD along the flow. By invoking ideas from non-equilibrium statistical mechanics we derive an efficient training procedure by which both the sampler's and the flow's parameters can be optimized end-to-end, and by which we can compute exact importance weights without having to marginalize out the randomness of the stochastic blocks. We illustrate the representational power, sampling efficiency and asymptotic correctness of SNFs on several benchmarks including applications to sampling molecular systems in equilibrium.

**********

GPU-Accelerated Primal Learning for Extremely Fast Large-Scale Classification
John T. Halloran, David M. Rocke

One of the most efficient methods to solve L2 -regularized primal problems, such as logistic regression and linear support vector machine (SVM) classification, is the widely used trust region Newton algorithm, TRON. While TRON has recently been shown to enjoy substantial speedups on shared-memory multi-core systems, exploiting graphical processing units (GPUs) to speed up the method is significantly more difficult, owing to the highly complex and heavily sequential nature of the algorithm. In this work, we show that using judicious GPU-optimization principles, TRON training time for different losses and feature representations may be drastically reduced. For sparse feature sets, we show that using GPUs to train logistic regression classifiers in LIBLINEAR is up to an order-of-magnitude faster than solely using multithreading. For dense feature sets–which impose far more stringent memory constraints–we show that GPUs substantially reduce the lengthy SVM learning times required for state-of-the-art proteomics analysis, leading to dramatic improvements over recently proposed speedups. Furthermore, we show how GPU speedups may be mixed with multithreading to enable such speedups when the dataset is too large for GPU memory requirements; on a massive dense proteomics dataset of nearly a quarter-billion data instances, these mixed-architecture speedups reduce SVM analysis time from over half a week to less than a single day while using limited GPU memory.

**********

Random Reshuffling is Not Always Better
Christopher M. De Sa

Many learning algorithms, such as stochastic gradient descent, are affected by the order in which training examples are used. It is often observed that sampling the training examples without-replacement, also known as random reshuffling, causes learning algorithms to converge faster. We give a counterexample to the Operator Inequality of Noncommutative Arithmetic and Geometric Means, a longstanding conjecture that relates to the performance of random reshuffling in learning algorithms (Recht and Ré, "Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences," COLT 2012). We use this to give an example of a learning task and algorithm for which with-replacement random sampling actually outperforms random reshuffling.

**********

Model Agnostic Multilevel Explanations
Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, Amit Dhurandhar

In recent years, post-hoc local instance-level and global dataset-level explainability of black-box models has received a lot of attention. Lesser attention has been given to obtaining insights at intermediate or group levels, which is a need outlined in recent works that study the challenges in realizing the guidelines in the General Data Protection Regulation (GDPR). In this paper, we propose a meta-method that, given a typical local explainability method, can build a multilevel explanation tree. The leaves of this tree correspond to local explanations, the root corresponds to global explanation, and intermediate levels correspond to explanations for groups of data points that it automatically clusters. The method can also leverage side information, where users can specify points for which they may want the explanations to be similar. We argue that such a multilevel

structure can also be an effective form of communication, where one could obtai n few explanations that characterize the entire dataset by considering an approp riate level in our explanation tree. Explanations for novel test points can be c ost-efficiently obtained by associating them with the closest training points. W hen the local explainability technique is generalized additive (viz. LIME, GAMs) , we develop fast approximate algorithm for building the multilevel tree and stu dy its convergence behavior. We show that we produce high fidelity sparse explan ations on several public datasets and also validate the effectiveness of the pro posed technique based on two human studies -- one with experts and the other wit h non-expert users -- on real world datasets.
**********

NeuMiss networks: differentiable programming for supervised learning with missin g values.
Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, Gael Varoquaux
The presence of missing values makes supervised learning much more challenging. Indeed, previous work has shown that even when the response is a linear function of the complete data, the optimal predictor is a complex function of the observ ed entries and the missingness indicator. As a result, the computational or samp le complexities of consistent approaches depend on the number of missing pattern s, which can be exponential in the number of dimensions. In this work, we derive the analytical form of the optimal predictor under a linearity assumption and v arious missing data mechanisms including Missing at Random (MAR) and self-maskin g (Missing Not At Random). Based on a Neumann-series approximation of the optima l predictor, we propose a new principled architecture, named NeuMiss networks. T heir originality and strength come from the use of a new type of non-linearity: the multiplication by the missingness indicator. We provide an upper bound on th e Bayes risk of NeuMiss networks, and show that they have good predictive accura cy with both a number of parameters and a computational complexity independent o f the number of missing data patterns. As a result they scale well to problems w ith many features, and remain statistically efficient for medium-sized samples. Moreover, we show that, contrary to procedures using EM or imputation, they are robust to the missing data mechanism, including difficult MNAR settings such as self-masking.
**********

Revisiting Parameter Sharing for Automatic Neural Channel Number Search
Jiaxing Wang, Haoli Bai, Jiaxiang Wu, Xupeng Shi, Junzhou Huang, Irwin King, Mic hael Lyu, Jian Cheng
Recent advances in neural architecture search inspire many channel number search algorithms~(CNS) for convolutional neural networks. To improve searching effici ency, parameter sharing is widely applied, which reuses parameters among differe nt channel configurations. Nevertheless, it is unclear how parameter sharing aff ects the searching process. In this paper, we aim at providing a better understa nding and exploitation of parameter sharing for CNS. Specifically, we propose af fine parameter sharing~(APS) as a general formulation to unify and quantitativel y analyze existing channel search algorithms. It is found that with parameter sh aring, weight updates of one architecture can simultaneously benefit other candi dates. However, it also results in less confidence in choosing good architecture s. We thus propose a new strategy of parameter sharing towards a better balance between training efficiency and architecture discrimination. Extensive analysis and experiments demonstrate the superiority of the proposed strategy in channel configuration against many state-of-the-art counterparts on benchmark datasets.
**********

Differentially-Private Federated Linear Bandits
Abhimanyu Dubey, Alex `Sandy' Pentland
The rapid proliferation of decentralized learning systems mandates the need for differentially-private cooperative learning. In this paper, we study this in co ntext of the contextual linear bandit: we consider a collection of agents cooper ating to solve a common contextual bandit, while ensuring that their communicati on remains private. For this problem, we devise FedUCB, a multiagent private alg orithm for both centralized and decentralized (peer-to-peer) federated learning.

We provide a rigorous technical analysis of its utility in terms of regret, imp
roving several results in cooperative bandit learning, and provide rigorous priv
acy guarantees as well. Our algorithms provide competitive performance both in t
erms of pseudoregret bounds and empirical benchmark performance in various multi
-agent settings.
**********
Is Plug-in Solver Sample-Efficient for Feature-based Reinforcement Learning?
Qiwen Cui, Lin 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.
**********
Learning Physical Graph Representations from Visual Scenes
Daniel Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jer
emy Schwartz, Li F. Fei-Fei, Jiajun Wu, Josh Tenenbaum, Daniel L. Yamins
Convolutional Neural Networks (CNNs) have proved exceptional at learning represe
ntations for visual object categorization. However, CNNs do not explicitly encod
e objects, parts, and their physical properties, which has limited CNNs' success
 on tasks that require structured understanding of visual scenes. To overcome th
ese limitations, we introduce the idea of ``Physical Scene Graphs'' (PSGs), whic
h represent scenes as hierarchical graphs, with nodes in the hierarchy correspon
ding intuitively to object parts at different scales, and edges to physical conn
ections between parts. Bound to each node is a vector of latent attributes that
intuitively represent object properties such as surface shape and texture.
We also describe PSGNet, a network architecture that learns to extract PSGs by r
econstructing scenes through a PSG-structured bottleneck. PSGNet augments standa
rd CNNs by including: recurrent feedback connections to combine low and high-lev
el image information; graph pooling and vectorization operations that convert sp
atially-uniform feature maps into object-centric graph structures; and perceptua
l grouping principles to encourage the identification of meaningful scene elemen
ts. We show that PSGNet outperforms alternative self-supervised scene representa
tion algorithms at scene segmentation tasks, especially on complex real-world im
ages, and generalizes well to unseen object types and scene arrangements. PSGNet
 is also able learn from physical motion, enhancing scene estimates even for sta
tic images. We present a series of ablation studies illustrating the importance
of each component of the PSGNet architecture, analyses showing that learned late
nt attributes capture intuitive scene properties, and illustrate the use of PSGs
 for compositional scene inference.
**********
Deep Graph Pose: a semi-supervised deep graphical model for improved animal pose
 tracking
Anqi Wu, Estefany Kelly Buchanan, Matthew Whiteway, Michael Schartner, Guido Mei
jer, Jean-Paul Noel, Erica Rodriguez, Claire Everett, Amy Norovich, Evan Schaffe
r, Neeli Mishra, C. Daniel Salzman, Dora Angelaki, Andrés Bendesky, The Internat
ional Brain Laboratory The International Brain Laboratory, John P. Cunningham, L
iam Paninski
Noninvasive behavioral tracking of animals is crucial for many scientific invest
igations. Recent transfer learning approaches for behavioral tracking have consi
derably advanced the state of the art. Typically these methods treat each video
frame and each object to be tracked independently. In this work, we improve on t
hese methods (particularly in the regime of few training labels) by leveraging
the rich spatiotemporal structures pervasive in behavioral video --- specificall
y, the spatial statistics imposed by physical constraints (e.g., paw to elbow di
stance), and the temporal statistics imposed by smoothness from frame to frame.
We propose a probabilistic graphical model built on top of deep neural networks,
 Deep Graph Pose (DGP), to leverage these useful spatial and temporal constraint
s, and develop an efficient structured variational approach to perform inference
 in this model.  The resulting semi-supervised model exploits both labeled and u
nlabeled frames to achieve significantly more accurate and robust tracking while

requiring users to label fewer training frames. In turn, these tracking improvements enhance performance on downstream applications, including robust unsupervised segmentation of behavioral syllables,'' and estimation of interpretabledisentangled'' low-dimensional representations of the full behavioral video. Open source code is available at \href{\CodeLink}{https://github.com/paninski-lab/deepgraphpose}.
**********

Meta-learning from Tasks with Heterogeneous Attribute Spaces
Tomoharu Iwata, Atsutoshi Kumagai
We propose a heterogeneous meta-learning method that trains a model on tasks with various attribute spaces, such that it can solve unseen tasks whose attribute spaces are different from the training tasks given a few labeled instances. Although many meta-learning methods have been proposed, they assume that all training and target tasks share the same attribute space, and they are inapplicable when attribute sizes are different across tasks. Our model infers latent representations of each attribute and each response from a few labeled instances using an inference network. Then, responses of unlabeled instances are predicted with the inferred representations using a prediction network. The attribute and response representations enable us to make predictions based on the task-specific properties of attributes and responses even when attribute and response sizes are different across tasks. In our experiments with synthetic datasets and 59 datasets in OpenML, we demonstrate that our proposed method can predict the responses given a few labeled instances in new tasks after being trained with tasks with heterogeneous attribute spaces.
**********

Estimating decision tree learnability with polylogarithmic sample complexity
Guy Blanc, Neha Gupta, Jane Lange, Li-Yang Tan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Sparse Symplectically Integrated Neural Networks
Daniel DiPietro, Shiying Xiong, Bo Zhu
We introduce Sparse Symplectically Integrated Neural Networks (SSINNs), a novel model for learning Hamiltonian dynamical systems from data. SSINNs combine fourth-order symplectic integration with a learned parameterization of the Hamiltonian obtained using sparse regression through a mathematically elegant function space. This allows for interpretable models that incorporate symplectic inductive biases and have low memory requirements. We evaluate SSINNs on four classical Hamiltonian dynamical problems: the Hénon-Heiles system, nonlinearly coupled oscillators, a multi-particle mass-spring system, and a pendulum system. Our results demonstrate promise in both system prediction and conservation of energy, often outperforming the current state-of-the-art black-box prediction techniques by an order of magnitude. Further, SSINNs successfully converge to true governing equations from highly limited and noisy data, demonstrating potential applicability in the discovery of new physical governing equations.
**********

Continuous Object Representation Networks: Novel View Synthesis without Target View Supervision
Nicolai Hani, Selim Engin, Jun-Jee Chao, Volkan Isler
Novel View Synthesis (NVS) is concerned with synthesizing views under camera viewpoint transformations from one or multiple input images. NVS requires explicit reasoning about 3D object structure and unseen parts of the scene to synthesize convincing results. As a result, current approaches typically rely on supervised training with either ground truth 3D models or multiple target images. We propose Continuous Object Representation Networks (CORN), a conditional architecture that encodes an input image's geometry and appearance that map to a 3D consistent scene representation. We can train CORN with only two source images per object by combining our model with a neural renderer. A key feature of CORN is that it

requires no ground truth 3D models or target view supervision. Regardless, CORN performs well on challenging tasks such as novel view synthesis and single-view 3D reconstruction and achieves performance comparable to state-of-the-art approaches that use direct supervision. For up-to-date information, data, and code, please see our project page: https://nicolaihaeni.github.io/corn/.
**********
Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence
Thomas Sutter, Imant Daunhawer, Julia Vogt
Learning from different data types is a long-standing goal in machine learning research, as multiple information sources co-occur when describing natural phenomena. However, existing generative models that approximate a multimodal ELBO rely on difficult or inefficient training schemes to learn a joint distribution and the dependencies between modalities.
In this work, we propose a novel, efficient objective function that utilizes the Jensen-Shannon divergence for multiple distributions. It simultaneously approximates the unimodal and joint multimodal posteriors directly via a dynamic prior. In addition, we theoretically prove that the new multimodal JS-divergence (mmJSD) objective optimizes an ELBO.
In extensive experiments, we demonstrate the advantage of the proposed mmJSD model compared to previous work in unsupervised, generative learning tasks.
**********
Solver-in-the-Loop: Learning from Differentiable Physics to Interact with Iterative PDE-Solvers
Kiwon Um, Robert Brand, Yun (Raymond) Fei, Philipp Holl, Nils Thuerey
Finding accurate solutions to partial differential equations (PDEs) is a crucial task in all scientific and engineering disciplines. It has recently been shown that machine learning methods can improve the solution accuracy by correcting for effects not captured by the discretized PDE. We target the problem of reducing numerical errors of iterative PDE solvers and compare different learning approaches for finding complex correction functions. We find that previously used learning approaches are significantly outperformed by methods that integrate the solver into the training loop and thereby allow the model to interact with the PDE during training. This provides the model with realistic input distributions that take previous corrections into account, yielding improvements in accuracy with stable rollouts of several hundred recurrent evaluation steps and surpassing even tailored supervised variants. We highlight the performance of the differentiable physics networks for a wide variety of PDEs, from non-linear advection-diffusion systems to three-dimensional Navier-Stokes flows.
**********
Reinforcement Learning with General Value Function Approximation: Provably Efficient Approach via Bounded Eluder Dimension
Ruosong Wang, Russ R. Salakhutdinov, Lin 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.
**********
Predicting Training Time Without Training
Luca Zancato, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika, Stefano Soatto
We tackle the problem of predicting the number of optimization steps that a pre-trained deep network needs to converge to a given value of the loss function. To do so, we leverage the fact that the training dynamics of a deep network during fine-tuning are well approximated by those of a linearized model. This allows us to approximate the training loss and accuracy at any point during training by solving a low-dimensional Stochastic Differential Equation (SDE) in function space. Using this result, we are able to predict the time it takes for Stochastic Gradient Descent (SGD) to fine-tune a model to a given loss without having to perform any training.
In our experiments, we are able to predict training time of a ResNet within a 20

\% error margin on a variety of datasets and hyper-parameters, at a 30 to 45-fold reduction in cost compared to actual training. We also discuss how to further reduce the computational and memory cost of our method, and in particular we show that by exploiting the spectral properties of the gradients' matrix it is possible to predict training time on a large dataset while processing only a subset of the samples.

**********

How does This Interaction Affect Me? Interpretable Attribution for Feature Interactions

Michael Tsang, Sirisha Rambhatla, Yan Liu

Machine learning transparency calls for interpretable explanations of how inputs relate to predictions. Feature attribution is a way to analyze the impact of features on predictions. Feature interactions are the contextual dependence between features that jointly impact predictions. There are a number of methods that extract feature interactions in prediction models; however, the methods that assign attributions to interactions are either uninterpretable, model-specific, or non-axiomatic. We propose an interaction attribution and detection framework called Archipelago which addresses these problems and is also scalable in real-world settings. Our experiments on standard annotation labels indicate our approach provides significantly more interpretable explanations than comparable methods, which is important for analyzing the impact of interactions on predictions. We also provide accompanying visualizations of our approach that give new insights into deep neural networks.

**********

Optimal Adaptive Electrode Selection to Maximize Simultaneously Recorded Neuron Yield

John Choi, Krishan Kumar, Mohammad Khazali, Katie Wingel, Mahdi Choudhury, Adam S. Charles, Bijan Pesaran

Neural-Matrix style, high-density electrode arrays for brain-machine interfaces (BMIs) and neuroscientific research require the use of multiplexing: Each recording channel can be routed to one of several electrode sites on the array. This capability allows the user to flexibly distribute recording channels to the locations where the most desirable neural signals can be resolved. For example, in the Neuropixel probe, 960 electrodes can be addressed by 384 recording channels. However, currently no adaptive methods exist to use recorded neural data to optimize/customize the electrode selections per recording context. Here, we present an algorithm called classification-based selection (CBS) that optimizes the joint electrode selections for all recording channels so as to maximize isolation quality of detected neurons. We show, in experiments using Neuropixels in non-human primates, that this algorithm yields a similar number of isolated neurons as would be obtained if all electrodes were recorded simultaneously. Neuron counts were 41-85% improved over previously published electrode selection strategies. The neurons isolated from electrodes selected by CBS were a 73% match, by spike timing, to the complete set of recordable neurons around the probe. The electrodes selected by CBS exhibited higher average per-recording-channel signal-to-noise ratio. CBS, and selection optimization in general, could play an important role in development of neurotechnologies for BMI, as signal bandwidth becomes an increasingly limiting factor. Code and experimental data have been made available.

**********

Neurosymbolic Reinforcement Learning with Formally Verified Exploration

Greg Anderson, Abhinav Verma, Isil Dillig, Swarat Chaudhuri

We present REVEL, a partially neural reinforcement learning (RL) framework for provably safe exploration in continuous state and action spaces. A key challenge for provably safe deep RL is that repeatedly verifying neural networks within a learning loop is computationally infeasible. We address this challenge using two policy classes: a general, neurosymbolic class with approximate gradients and a more restricted class of symbolic policies that allows efficient verification. Our learning algorithm is a mirror descent over policies: in each iteration, it safely lifts a symbolic policy into the neurosymbolic space, performs safe gradient updates to the resulting policy, and projects the updated policy into the s

afe symbolic subset, all without requiring explicit verification of neural netwo
rks. Our empirical results show that REVEL enforces safe exploration in many sce
narios in which Constrained Policy Optimization does not, and that it can discov
er policies that outperform those learned through prior approaches to verified e
xploration.
**********

Wavelet Flow: Fast Training of High Resolution Normalizing Flows
Jason J. Yu, Konstantinos G. Derpanis, Marcus A. Brubaker
Normalizing flows are a class of probabilistic generative models which allow for
 both fast density computation and efficient sampling and are effective at model
ling complex distributions like images. A drawback among current methods is thei
r significant training cost, sometimes requiring months of GPU training time to
achieve state-of-the-art results. This paper introduces Wavelet Flow, a multi-sc
ale, normalizing flow architecture based on wavelets. A Wavelet Flow has an expl
icit representation of signal scale that inherently includes models of lower res
olution signals and conditional generation of higher resolution signals, i.e., s
uper resolution. A major advantage of Wavelet Flow is the ability to construct g
enerative models for high resolution data (e.g., 1024 × 1024 images) that are im
practical with previous models. Furthermore, Wavelet Flow is competitive with pr
evious normalizing flows in terms of bits per dimension on standard (low resolut
ion) benchmarks while being up to 15× faster to train.
**********

Multi-task Batch Reinforcement Learning with Metric Learning
Jiachen Li, Quan Vuong, Shuang Liu, Minghua Liu, Kamil Ciosek, Henrik Christense
n, Hao Su
We tackle the Multi-task Batch Reinforcement Learning problem. Given multiple da
tasets collected from different tasks, we train a multi-task policy to perform w
ell in unseen tasks sampled from the same distribution. The task identities of t
he unseen tasks are not provided. To perform well, the policy must infer the tas
k identity from collected transitions by modelling its dependency on states, act
ions and rewards. Because the different datasets may have state-action distribut
ions with large divergence, the task inference module can learn to ignore the re
wards and spuriously correlate \textit{only} state-action pairs to the task iden
tity, leading to poor test time performance. To robustify task inference, we pro
pose a novel application of the triplet loss. To mine hard negative examples, we
 relabel the transitions from the training tasks by approximating their reward f
unctions. When we allow further training on the unseen tasks, using the trained
policy as an initialization leads to significantly faster convergence compared t
o randomly initialized policies (up to 80% improvement and across 5 different Mu
joco task distributions). We name our method \textbf{MBML} (\textbf{M}ulti-task
\textbf{B}atch RL with \textbf{M}etric \textbf{L}earning).
**********

On 1/n neural representation and robustness
Josue Nassar, Piotr Sokol, Sueyeon Chung, Kenneth D. Harris, Il Memming Park
Understanding the nature of representation in neural networks is a goal shared b
y neuroscience and machine learning.
It is therefore exciting that both fields converge not only on shared questions
but also on similar approaches.
A pressing question in these areas is understanding how the structure of the rep
resentation used by neural networks affects both their generalization, and robus
tness to perturbations.
In this work, we investigate the latter by juxtaposing experimental results rega
rding the covariance spectrum of neural representations in the mouse V1 (Stringe
r et al) with artificial neural networks.
We use adversarial robustness to probe Stringer et al’s theory regarding the cau
sal role of a 1/n covariance spectrum.
We empirically investigate the benefits such a neural code confers in neural net
works, and illuminate its role in multi-layer architectures.
Our results show that imposing the experimentally observed structure on artifici
al neural networks makes them more robust to adversarial attacks.

Moreover, our findings complement the existing theory relating wide neural netwo
rks to kernel methods, by showing the role of intermediate representations.
**********

Boundary thickness and robustness in learning models
Yaoqing Yang, Rajiv Khanna, Yaodong Yu, Amir Gholami, Kurt Keutzer, Joseph E. Go
nzalez, Kannan Ramchandran, Michael W. Mahoney
Robustness of machine learning models to various adversarial and non-adversarial
 corruptions continues to be of interest. In this paper, we introduce the notion
 of the boundary thickness of a classifier, and we describe its connection with
and usefulness for model robustness. Thick decision boundaries lead to improved
performance, while thin decision boundaries lead to overfitting (e.g., measured
by the robust generalization gap between training and testing) and lower robustn
ess. We show that a thicker boundary helps improve robustness against adversaria
l examples (e.g., improving the robust test accuracy of adversarial training), a
s well as so-called out-of-distribution (OOD) transforms, and we show that many
commonly-used regularization and data augmentation procedures can increase bound
ary thickness. On the theoretical side, we establish that maximizing boundary th
ickness is akin to minimizing the so-called mixup loss. Using these observations
, we can show that noise-augmentation on mixup training further increases bounda
ry thickness, thereby combating vulnerability to various forms of adversarial at
tacks and OOD transforms. We can also show that the performance improvement in s
everal recent lines of work happens in conjunction with a thicker boundary.
**********

Demixed shared component analysis of neural population data from multiple brain
areas
Yu Takagi, Steven Kennerley, Jun-ichiro Hirayama, Laurence Hunt
Recent advances in neuroscience data acquisition allow for the simultaneous reco
rding of large populations of neurons across multiple brain areas while subjects
 perform complex cognitive tasks. Interpreting these data requires us to index h
ow task-relevant information is shared across brain regions, but this is often c
onfounded by the mixing of different task parameters at the single neuron level.
 Here, inspired by a method developed for a single brain area, we introduce a ne
w technique for demixing variables across multiple brain areas, called demixed s
hared component analysis (dSCA). dSCA decomposes population activity into a few
components, such that the shared components capture the maximum amount of shared
 information across brain regions while also depending on relevant task paramete
rs. This yields interpretable components that express which variables are shared
 between different brain regions and when this information is shared across time
. To illustrate our method, we reanalyze two datasets recorded during decision-m
aking tasks in rodents and macaques. We find that dSCA provides new insights int
o the shared computation between different brain areas in these datasets, relati
ng to several different aspects of decision formation.
**********

Learning Kernel Tests Without Data Splitting
Jonas Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, Krikamol Muandet
Modern large-scale kernel-based tests such as maximum mean discrepancy (MMD) and
 kernelized Stein discrepancy (KSD) optimize kernel hyperparameters on a held-ou
t sample via data splitting to obtain the most powerful test statistics. While d
ata splitting results in a tractable null distribution, it suffers from a reduct
ion in test power due to smaller test sample size. Inspired by the selective inf
erence framework, we propose an approach that enables learning the hyperparamete
rs and testing on the full sample without data splitting. Our approach can corre
ctly calibrate the test in the presence of such dependency, and yield a test thr
eshold in closed form. At the same significance level, our approach’s test power
 is empirically larger than that of the data-splitting approach, regardless of i
ts split proportion.
**********

Unsupervised Data Augmentation for Consistency Training
Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, Quoc Le
Semi-supervised learning lately has shown much promise in improving deep learnin

g models when labeled data is scarce. Common among recent approaches is the use of consistency training on a large amount of unlabeled data to constrain model predictions to be invariant to input noise. In this work, we present a new perspective on how to effectively noise unlabeled examples and argue that the quality of noising, specifically those produced by advanced data augmentation methods, plays a crucial role in semi-supervised learning. By substituting simple noising operations with advanced data augmentation methods such as RandAugment and back-translation, our method brings substantial improvements across six language and three vision tasks under the same consistency training framework. On the IMDb text classification dataset, with only 20 labeled examples, our method achieves an error rate of 4.20, outperforming the state-of-the-art model trained on 25,000 labeled examples. On a standard semi-supervised learning benchmark, CIFAR-10, our method outperforms all previous approaches and achieves an error rate of 5.43 with only 250 examples. Our method also combines well with transfer learning, e.g., when finetuning from BERT, and yields improvements in high-data regime, such as ImageNet, whether when there is only 10% labeled data or when a full labeled set with 1.3M extra unlabeled examples is used. Code is available at https://github.com/google-research/uda.

**********

Subgroup-based Rank-1 Lattice Quasi-Monte Carlo

Yueming LYU, Yuan Yuan, Ivor Tsang

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

**********

Minibatch vs Local SGD for Heterogeneous Distributed Learning

Blake E. Woodworth, Kumar Kshitij Patel, Nati Srebro

We analyze Local SGD (aka parallel or federated SGD) and Minibatch SGD in the heterogeneous distributed setting, where each machine has access to stochastic gradient estimates for a different, machine-specific, convex objective; the goal is to optimize w.r.t.~the average objective; and machines can only communicate intermittently.  We argue that, (i) Minibatch SGD (even without acceleration) dominates all existing analysis of Local SGD in this setting, (ii) accelerated Minibatch SGD is optimal when the heterogeneity is high, and (iii) present the first upper bound for Local SGD that improves over Minibatch SGD in a non-homogeneous regime.

**********

Multi-task Causal Learning with Gaussian Processes

Virginia Aglietti, Theodoros Damoulas, Mauricio Álvarez, Javier González

This paper studies the problem of learning the correlation structure of a set of intervention functions defined on the directed acyclic graph (DAG) of a causal model. This is useful when we are interested in jointly learning the causal effects of interventions on different subsets of variables in a DAG, which is common in field such as healthcare or operations research. We propose the first multi-task causal Gaussian process (GP) model, which we call DAG-GP, that allows for information sharing across continuous interventions and across experiments on different variables. DAG-GP accommodates different assumptions in terms of data availability and captures the correlation between functions lying in input spaces of different dimensionality via a well-defined integral operator. We give theoretical results detailing when and how the DAG-GP model can be formulated depending on the DAG. We test both the quality of its predictions and its calibrated uncertainties. Compared to single-task models, DAG-GP achieves the best fitting performance in a variety of real and synthetic settings. In addition, it helps to select optimal interventions faster than competing approaches when used within sequential decision making frameworks, like active learning or Bayesian optimization.

**********

Proximity Operator of the Matrix Perspective Function and its Applications

Joong-Ho (Johann) Won

We show that the matrix perspective function, which is jointly convex in the Cartesian product of a standard Euclidean vector space and a conformal space of symmetric matrices, has a proximity operator in an almost closed form. The only implicit part is to solve a semismooth, univariate root finding problem. We uncover the connection between our problem of study and the matrix nearness problem. Through this connection, we propose a quadratically convergent Newton algorithm for the root finding problem.Experiments verify that the evaluation of the proximity operator requires at most 8 Newton steps, taking less than 5s for 2000 by 2000 matrices on a standard laptop. Using this routine as a building block, we demonstrate the usefulness of the studied proximity operator in constrained maximum likelihood estimation of Gaussian mean and covariance, peudolikelihood-based graphical model selection, and a matrix variant of the scaled lasso problem.
**********
Generative 3D Part Assembly via Dynamic Graph Learning
jialei huang, Guanqi Zhan, Qingnan Fan, Kaichun Mo, Lin Shao, Baoquan Chen, Leonidas J. Guibas, Hao Dong
Autonomous part assembly is a challenging yet crucial task in 3D computer vision and robotics. Analogous to buying an IKEA furniture, given a set of 3D parts that can assemble a single shape, an intelligent agent needs to perceive the 3D part geometry, reason to propose pose estimations for the input parts, and finally call robotic planning and control routines for actuation. In this paper, we focus on the pose estimation subproblem from the vision side involving geometric and relational reasoning over the input part geometry. Essentially, the task of generative 3D part assembly is to predict a 6-DoF part pose, including a rigid rotation and translation, for each input part that assembles a single 3D shape as the final output. To tackle this problem, we propose an assembly-oriented dynamic graph learning framework that leverages an iterative graph neural network as a backbone. It explicitly conducts sequential part assembly refinements in a coarse-to-fine manner, exploits a pair of part relation reasoning module and part aggregation module for dynamically adjusting both part features and their relations in the part graph. We conduct extensive experiments and quantitative comparisons to three strong baseline methods, demonstrating the effectiveness of the proposed approach.
**********
Improving Natural Language Processing Tasks with Human Gaze-Guided Neural Attention
Ekta Sood, Simon Tannert, Philipp Mueller, Andreas Bulling
A lack of corpora has so far limited advances in integrating human gaze data as a
supervisory signal in neural attention mechanisms for natural language processing
(NLP). We propose a novel hybrid text saliency model (TSM) that, for the first
time, combines a cognitive model of reading with explicit human gaze supervision
in a single machine learning framework. On four different corpora we demonstrate
that our hybrid TSM duration predictions are highly correlated with human gaze
ground truth. We further propose a novel joint modeling approach to integrate TSM
predictions into the attention layer of a network designed for a specific upstream
NLP task without the need for any task-specific human gaze data. We demonstrate
that our joint model outperforms the state of the art in paraphrase generation on
the Quora Question Pairs corpus by more than 10% in BLEU-4 and achieves
state of the art performance for sentence compression on the challenging Google
Sentence Compression corpus. As such, our work introduces a practical approach
for bridging between data-driven and cognitive models and demonstrates a new
way to integrate human gaze-guided neural attention into NLP tasks.
**********
The Power of Comparisons for Actively Learning Linear Classifiers
Max Hopkins, Daniel Kane, Shachar Lovett

In the world of big data, large but costly to label datasets dominate many fields. Active learning, a semi-supervised alternative to the standard PAC-learning model, was introduced to explore whether adaptive labeling could learn concepts with exponentially fewer labeled samples. While previous results show that active learning performs no better than its supervised alternative for important concept classes such as linear separators, we show that by adding weak distributional assumptions and allowing comparison queries, active learning requires exponentially fewer samples. Further, we show that these results hold as well for a stronger model of learning called Reliable and Probably Useful (RPU) learning. In this model, our learner is not allowed to make mistakes, but may instead answer ``I don't know.'' While previous negative results showed this model to have intractably large sample complexity for label queries, we show that comparison queries make RPU-learning at worst logarithmically more expensive in both the passive and active regimes.
**********

From Boltzmann Machines to Neural Networks and Back Again
Surbhi Goel, Adam Klivans, Frederic Koehler
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Crush Optimism with Pessimism: Structured Bandits Beyond Asymptotic Optimality
Kwang-Sung Jun, Chicheng 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.
**********

Pruning neural networks without any data by iteratively conserving synaptic flow
Hidenori Tanaka, Daniel Kunin, Daniel L. Yamins, Surya Ganguli
Pruning the parameters of deep neural networks has generated intense interest due to potential savings in time, memory and energy both during training and at test time. Recent works have identified, through an expensive sequence of training and pruning cycles, the existence of winning lottery tickets or sparse trainable subnetworks at initialization. This raises a foundational question: can we identify highly sparse trainable subnetworks at initialization, without ever training, or indeed without ever looking at the data? We provide an affirmative answer to this question through theory driven algorithm design. We first mathematically formulate and experimentally verify a conservation law that explains why existing gradient-based pruning algorithms at initialization suffer from layer-collapse, the premature pruning of an entire layer rendering a network untrainable. This theory also elucidates how layer-collapse can be entirely avoided, motivating a novel pruning algorithm Iterative Synaptic Flow Pruning (SynFlow). This algorithm can be interpreted as preserving the total flow of synaptic strengths through the network at initialization subject to a sparsity constraint. Notably, this algorithm makes no reference to the training data and consistently competes with or outperforms existing state-of-the-art pruning algorithms at initialization over a range of models (VGG and ResNet), datasets (CIFAR-10/100 and Tiny ImageNet), and sparsity constraints (up to 99.99 percent). Thus our data-agnostic pruning algorithm challenges the existing paradigm that, at initialization, data must be used to quantify which synapses are important.
**********

Detecting Interactions from Neural Networks via Topological Analysis
Zirui Liu, Qingquan Song, Kaixiong Zhou, Ting-Hsiang Wang, Ying Shan, Xia Hu
Detecting statistical interactions between input features is a crucial and challenging task. Recent advances demonstrate that it is possible to extract learned interactions from trained neural networks. It has also been observed that, in neural networks, any interacting features must follow a strongly weighted connection to common hidden units. Motivated by the observation, in this paper, we propo

se to investigate the interaction detection problem from a novel topological per spective by analyzing the connectivity in neural networks. Specially, we propose a new measure for quantifying interaction strength, based upon the well-receive d theory of persistent homology. Based on this measure, a Persistence Interactio n Dection (PID) algorithm is developed to efficiently detect interactions. Our p roposed algorithm is evaluated across a number of interaction detection tasks on several synthetic and real-world datasets with different hyperparameters. Exper imental results validate that the PID algorithm outperforms the state-of-the-art baselines.

**********

Neural Bridge Sampling for Evaluating Safety-Critical Autonomous Systems

Aman Sinha, Matthew O'Kelly, Russ Tedrake, John C. Duchi

Learning-based methodologies increasingly find applications in safety-critical d omains like autonomous driving and medical robotics. Due to the rare nature of d angerous events, real-world testing is prohibitively expensive and unscalable. In this work, we employ a probabilistic approach to safety evaluation in simulat ion, where we are concerned with computing the probability of dangerous events. We develop a novel rare-event simulation method that combines exploration, explo itation, and optimization techniques to find failure modes and estimate their ra te of occurrence. We provide rigorous guarantees for the performance of our meth od in terms of both statistical and computational efficiency. Finally, we demons trate the efficacy of our approach on a variety of scenarios, illustrating its u sefulness as a tool for rapid sensitivity analysis and model comparison that are essential to developing and testing safety-critical autonomous systems.

**********

Interpretable and Personalized Apprenticeship Scheduling: Learning Interpretable Scheduling Policies from Heterogeneous User Demonstrations

Rohan Paleja, Andrew Silva, Letian Chen, Matthew Gombolay

Requests for name changes in the electronic proceedings will be accepted with 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.

**********

Task-Agnostic Online Reinforcement Learning with an Infinite Mixture of Gaussian Processes

Mengdi Xu, Wenhao Ding, Jiacheng Zhu, ZUXIN LIU, Baiming Chen, Ding Zhao

Continuously learning to solve unseen tasks with limited experience has been ext ensively pursued in meta-learning and continual learning, but with restricted as sumptions such as accessible task distributions, independently and identically d istributed tasks, and clear task delineations. However, real-world physical task s frequently violate these assumptions, resulting in performance degradation. Th is paper proposes a continual online model-based reinforcement learning approach that does not require pre-training to solve task-agnostic problems with unknown task boundaries. We maintain a mixture of experts to handle nonstationarity, an d represent each different type of dynamics with a Gaussian Process to efficient ly leverage collected data and expressively model uncertainty. We propose a tran sition prior to account for the temporal dependencies in streaming data and upda te the mixture online via sequential variational inference. Our approach reliabl y handles the task distribution shift by generating new models for never-before-seen dynamics and reusing old models for previously seen dynamics. In experiment s, our approach outperforms alternative methods in non-stationary tasks, includi ng classic control with changing dynamics and decision making in different drivi ng scenarios.

**********

Benchmarking Deep Learning Interpretability in Time Series Predictions

Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, Soheil Feizi

Saliency methods are used extensively to highlight the importance of input featu res in model predictions. These methods are mostly used in vision and language t asks, and their applications to time series data is relatively unexplored. In th is paper, we set out to extensively compare the performance of various saliency-

based interpretability methods across diverse neural architectures, including Re
current Neural Network, Temporal Convolutional Networks, and Transformers in a n
ew benchmark of synthetic time series data. We propose and report multiple metri
cs to empirically evaluate the performance of saliency methods for detecting fea
ture importance over time using both precision (i.e., whether identified feature
s contain meaningful signals) and recall (i.e., the number of features with sign
al identified as important). Through several experiments, we show that (i) in ge
neral, network architectures and saliency methods fail to reliably and accuratel
y identify feature importance over time in time series data, (ii) this failure i
s mainly due to the conflation of time and feature domains, and (iii) the qualit
y of saliency maps can be improved substantially by using our proposed two-step
temporal saliency rescaling (TSR) approach that first calculates the importance
of each time step before calculating the importance of each feature at a time st
ep.
**********

Federated Principal Component Analysis
Andreas Grammenos, Rodrigo Mendoza Smith, Jon Crowcroft, Cecilia Mascolo
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

(De)Randomized Smoothing for Certifiable Defense against Patch Attacks
Alexander Levine, Soheil Feizi
Patch adversarial attacks on images, in which the attacker can distort pixels wi
thin a region of bounded size, are an important threat model since they provide
a quantitative model for physical adversarial attacks. In this paper, we introdu
ce a certifiable defense against patch attacks that guarantees for a given image
 and patch attack size, no patch adversarial examples exist. Our method is relat
ed to the broad class of randomized smoothing robustness schemes which provide h
igh-confidence probabilistic robustness certificates. By exploiting the fact tha
t patch attacks are more constrained than general sparse attacks, we derive mean
ingfully large robustness certificates against them. Additionally, in contrast t
o smoothing-based defenses against L_p and sparse attacks, our defense method ag
ainst patch attacks is de-randomized, yielding improved, deterministic certifica
tes. Compared to the existing patch certification method proposed by (Chiang et
al., 2020), which relies on interval bound propagation, our method can be traine
d significantly faster, achieves high clean and certified robust accuracy on CIF
AR-10, and provides certificates at ImageNet scale. For example, for a 5-by-5 pa
tch attack on CIFAR-10, our method achieves up to around 57.6% certified accurac
y (with a classifier with around 83.8% clean accuracy), compared to at most 30.3
% certified accuracy for the existing method (with a classifier with around 47.8
% clean accuracy). Our results effectively establish a new state-of-the-art of c
ertifiable defense against patch attacks on CIFAR-10 and ImageNet.
**********

SMYRF - Efficient Attention using Asymmetric Clustering
Giannis Daras, Nikita Kitaev, Augustus Odena, 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.
**********

Introducing Routing Uncertainty in Capsule Networks
Fabio De Sousa Ribeiro, Georgios Leontidis, Stefanos Kollias
Rather than performing inefficient local iterative routing between adjacent caps
ule layers, we propose an alternative global view based on representing the inhe
rent uncertainty in part-object assignment. In our formulation, the local routin
g iterations are replaced with variational inference of part-object connections
in a probabilistic capsule network, leading to a significant speedup without sac
rificing performance. In this way, global context is also considered when routin

g capsules by introducing global latent variables that have direct influence on the objective function, and are updated discriminatively in accordance with the minimum description length (MDL) principle. We focus on enhancing capsule networ k properties, and perform a thorough evaluation on pose-aware tasks, observing i mprovements in performance over previous approaches whilst being more computatio nally efficient.
**********

A Simple and Efficient Smoothing Method for Faster Optimization and Local Explor ation
Kevin Scaman, Ludovic DOS SANTOS, Merwan Barlier, Igor Colin
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Hyperparameter Ensembles for Robustness and Uncertainty Quantification
Florian Wenzel, Jasper Snoek, Dustin Tran, Rodolphe Jenatton
Ensembles over neural network weights trained from different random initializati on, known as deep ensembles, achieve state-of-the-art accuracy and calibration. The recently introduced batch ensembles provide a drop-in replacement that is mo re parameter efficient. In this paper, we design ensembles not only over weights , but over hyperparameters to improve the state of the art in both settings. For best performance independent of budget, we propose hyper-deep ensembles, a simp le procedure that involves a random search over different hyperparameters, thems elves stratified across multiple random initializations. Its strong performance highlights the benefit of combining models with both weight and hyperparameter d iversity. We further propose a parameter efficient version, hyper-batch ensemble s, which builds on the layer structure of batch ensembles and self-tuning networ ks. The computational and memory costs of our method are notably lower than typi cal ensembles. On image classification tasks, with MLP, LeNet, ResNet 20 and Wid e ResNet 28-10 architectures, we improve upon both deep and batch ensembles.
**********

Neutralizing Self-Selection Bias in Sampling for Sortition
Bailey Flanigan, Paul Gölz, Anupam Gupta, Ariel D. Procaccia
Sortition is a political system in which decisions are made by panels of randoml y selected citizens. The process for selecting a sortition panel is traditionall y thought of as uniform sampling without replacement, which has strong fairness properties. In practice, however, sampling without replacement is not possible s ince only a fraction of agents is willing to participate in a panel when invited , and different demographic groups participate at different rates. In order to s till produce panels whose composition resembles that of the population, we devel op a sampling algorithm that restores close-to-equal representation probabilitie s for all agents while satisfying meaningful demographic quotas. As part of its input, our algorithm requires probabilities indicating how likely each volunteer in the pool was to participate. Since these participation probabilities are not directly observable, we show how to learn them, and demonstrate our approach us ing data on a real sortition panel combined with information on the general popu lation in the form of publicly available survey data.
**********

On the Convergence of Smooth Regularized Approximate Value Iteration Schemes
Elena Smirnova, Elvis Dohmatob
Entropy regularization, smoothing of Q-values and neural network function approx imator are key components of the state-of-the-art reinforcement learning (RL) al gorithms, such as Soft Actor-Critic~\cite{haarnoja2018soft}. Despite the widespr ead use, the impact of these core techniques on the convergence of RL algorithms is not yet fully understood. In this work, we analyse these techniques from err or propagation perspective using the approximate dynamic programming framework. In particular, our analysis shows that (1) value smoothing results in increased stability of the algorithm in exchange for slower convergence, (2) entropy regul arization reduces overestimation errors at the cost of modifying the original pr

oblem, (3) we study a combination of these techniques that describes the Soft Actor-Critic algorithm.
**********
Off-Policy Evaluation via the Regularized Lagrangian
Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, Dale Schuurmans
The recently proposed distribution correction estimation (DICE) family of estimators has advanced the state of the art in off-policy evaluation from behavior-agnostic data. While these estimators all perform some form of stationary distribution correction, they arise from different derivations and objective functions. In this paper, we unify these estimators as regularized Lagrangians of the same linear program. The unification allows us to expand the space of DICE estimators to new alternatives that demonstrate improved performance. More importantly, by analyzing the expanded space of estimators both mathematically and empirically we find that dual solutions offer greater flexibility in navigating the tradeoff between optimization stability and estimation bias, and generally provide superior estimates in practice.
**********
The LoCA Regret: A Consistent Metric to Evaluate Model-Based Behavior in Reinforcement Learning
Harm Van Seijen, Hadi Nekoei, Evan Racah, Sarath Chandar
Deep model-based Reinforcement Learning (RL) has the potential to substantially improve the sample-efficiency of deep RL. While various challenges have long held it back, a number of papers have recently come out reporting success with deep model-based methods. This is a great development, but the lack of a consistent metric to evaluate such methods makes it difficult to compare various approaches. For example, the common single-task sample-efficiency metric conflates improvements due to model-based learning with various other aspects, such as representation learning, making it difficult to assess true progress on model-based RL. To address this, we introduce an experimental setup to evaluate model-based behavior of RL methods, inspired by work from neuroscience on detecting model-based behavior in humans and animals. Our metric based on this setup, the Local Change Adaptation (LoCA) regret, measures how quickly an RL method adapts to a local change in the environment. Our metric can identify model-based behavior, even if the method uses a poor representation and provides insight in how close a method's behavior is from optimal model-based behavior. We use our setup to evaluate the model-based behavior of MuZero on a variation of the classic Mountain Car task.
**********
Neural Power Units
Niklas Heim, Tomas Pevny, Vasek Smidl
Conventional Neural Networks can approximate simple arithmetic operations, but fail to generalize beyond the range of numbers that were seen during training. Neural Arithmetic Units aim to overcome this difficulty, but current arithmetic units are either limited to operate on positive numbers or can only represent a subset of arithmetic operations. We introduce the Neural Power Unit (NPU) that operates on the full domain of real numbers and is capable of learning arbitrary power functions in a single layer. The NPU thus fixes the shortcomings of existing
arithmetic units and extends their expressivity. We achieve this by using complex
arithmetic without requiring a conversion of the network to complex numbers. A simplification of the unit to the RealNPU yields a highly transparent model. We show that the NPUs outperform their competitors in terms of accuracy and sparsity
on artificial arithmetic datasets, and that the RealNPU can discover the governing
equations of a dynamical system only from data.
**********
Towards Scalable Bayesian Learning of Causal DAGs
Jussi Viinikka, Antti Hyttinen, Johan Pensar, Mikko Koivisto
We give methods for Bayesian inference of directed acyclic graphs, DAGs, and the

induced causal effects from passively observed complete data. Our methods build on a recent Markov chain Monte Carlo scheme for learning Bayesian networks, which enables ef■cient approximate sampling from the graph posterior, provided that each node is assigned a small number K of candidate parents. We present algorithmic techniques to signi■cantly reduce the space and time requirements, which make the use of substantially larger values of K feasible. Furthermore, we investigate the problem of selecting the candidate parents per node so as to maximize the covered posterior mass. Finally, we combine our sampling method with a novel Bayesian approach for estimating causal effects in linear Gaussian DAG models. Numerical experiments demonstrate the performance of our methods in detecting ancestor–descendant relations, and in causal effect estimation our Bayesian method is shown to outperform previous approaches.
**********

A Dictionary Approach to Domain-Invariant Learning in Deep Networks
Ze Wang, Xiuyuan Cheng, Guillermo Sapiro, Qiang Qiu
In this paper, we consider domain-invariant deep learning by explicitly modeling domain shifts with only a small amount of domain-specific parameters in a Convolutional Neural Network (CNN).
By exploiting the observation that a convolutional filter can be well approximated as a linear combination of a small set of dictionary atoms, we show for the first time, both empirically and theoretically, that domain shifts can be effectively handled by decomposing a convolutional layer into a domain-specific atom layer and a domain-shared coefficient layer, while both remain convolutional.
An input channel will now first convolve spatially only with each respective domain-specific dictionary atom to ``absorb" domain variations, and then output channels are linearly combined using common decomposition coefficients trained to promote shared semantics across domains.
We use toy examples, rigorous analysis, and real-world examples with diverse datasets and architectures, to show the proposed plug-in framework's effectiveness in cross and joint domain performance and domain adaptation.
With the proposed architecture, we need only a small set of dictionary atoms to model each additional domain, which brings a negligible amount of additional parameters, typically a few hundred.
**********

Bootstrapping neural processes
Juho Lee, Yoonho Lee, Jungtaek Kim, Eunho Yang, Sung Ju Hwang, Yee Whye Teh
Unlike in the traditional statistical modeling for which a user typically hand-specify a prior, Neural Processes (NPs) implicitly define a broad class of stochastic processes with neural networks. Given a data stream, NP learns a stochastic process that best describes the data. While this ``data-driven'' way of learning stochastic processes has proven to handle various types of data, NPs still relies on an assumption that uncertainty in stochastic processes is modeled by a single latent variable, which potentially limits the flexibility. To this end, we propose the Bootstrapping Neural Process (BNP), a novel extension of the NP family using the bootstrap. The bootstrap is a classical data-driven technique for estimating uncertainty, which allows BNP to learn the stochasticity in NPs without assuming a particular form. We demonstrate the efficacy of BNP on various types of data and its robustness in the presence of model-data mismatch.
**********

Large-Scale Adversarial Training for Vision-and-Language Representation Learning
Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, Jingjing Liu
We present VILLA, the first known effort on large-scale adversarial training for vision-and-language (V+L) representation learning. VILLA consists of two training stages: (i) task-agnostic adversarial pre-training; followed by (ii) task-specific adversarial finetuning. Instead of adding adversarial perturbations on image pixels and textual tokens, we propose to perform adversarial training in the embedding space of each modality. To enable large-scale training, we adopt the ``free'' adversarial training strategy, and combine it with KL-divergence-based regularization to promote higher invariance in the embedding space. We apply VILLA to current best-performing V+L models, and achieve new state of the art on a w

ide range of tasks, including Visual Question Answering, Visual Commonsense Reas
oning, Image-Text Retrieval, Referring Expression Comprehension, Visual Entailme
nt, and NLVR2.
**********

Most ReLU Networks Suffer from $\ell^2$ Adversarial Perturbations
Amit Daniely, Hadas Shacham
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Compositional Visual Generation with Energy Based Models
Yilun Du, Shuang Li, Igor Mordatch
A vital aspect of human intelligence is the ability to compose increasingly comp
lex concepts out of simpler ideas, enabling both rapid learning and adaptation o
f knowledge. In this paper we show that energy-based models can exhibit this abi
lity by directly combining probability distributions. Samples from the combined
distribution correspond to compositions of concepts. For example, given a distri
bution for smiling faces, and another for male faces, we can combine them to gen
erate smiling male faces. This allows us to generate natural images that simulta
neously satisfy conjunctions, disjunctions, and negations of concepts. We evalua
te compositional generation abilities of our model on the CelebA dataset of natu
ral faces and synthetic 3D scene images. We also demonstrate other unique advant
ages of our model, such as the ability to continually learn and incorporate new
concepts, or infer compositions of concept properties underlying an image.
**********

Factor Graph Grammars
David Chiang, Darcey Riley
We propose the use of hyperedge replacement graph grammars for factor graphs, or
 factor graph grammars (FGGs) for short. FGGs generate sets of factor graphs and
 can describe a more general class of models than plate notation, dynamic graphi
cal models, case-factor diagrams, and sum-product networks can. Moreover, infere
nce can be done on FGGs without enumerating all the generated factor graphs. For
 finite variable domains (but possibly infinite sets of graphs), a generalizatio
n of variable elimination to FGGs allows exact and tractable inference in many s
ituations. For finite sets of graphs (but possibly infinite variable domains), a
 FGG can be converted to a single factor graph amenable to standard inference te
chniques.
**********

Erdos Goes Neural: an Unsupervised Learning Framework for Combinatorial Optimiza
tion on Graphs
Nikolaos Karalias, Andreas Loukas
Combinatorial optimization (CO) problems are notoriously challenging for neural
networks, especially in the absence of labeled instances. This work proposes an
unsupervised learning framework for CO problems on graphs that can provide integ
ral solutions of certified quality.
Inspired by Erdos' probabilistic method, we use a neural network to parametrize
a probability distribution over sets. Crucially, we show that when the network i
s optimized w.r.t. a suitably chosen loss, the learned distribution contains, wi
th controlled probability, a low-cost integral solution that obeys the constrain
ts of the combinatorial problem.
The probabilistic proof of existence is then derandomized to
decode the desired solutions. We demonstrate the efficacy of this approach to ob
tain valid
solutions to the maximum clique problem and to perform local graph clustering. O
ur method achieves competitive results on both real datasets and synthetic hard
instances.
**********

Autoregressive Score Matching
Chenlin Meng, Lantao Yu, Yang Song, Jiaming Song, Stefano Ermon

Autoregressive models use chain rule to define a joint probability distribution as a product of conditionals. These conditionals need to be normalized, imposing constraints on the functional families that can be used. To increase flexibility, we propose autoregressive conditional score models (AR-CSM) where we parameterize the joint distribution in terms of the derivatives of univariate log-conditionals (scores), which need not be normalized. To train AR-CSM, we introduce a new divergence between distributions named Composite Score Matching (CSM). For AR-CSM models, this divergence between data and model distributions can be computed and optimized efficiently, requiring no expensive sampling or adversarial training. Compared to previous score matching algorithms, our method is more scalable to high dimensional data and more stable to optimize. We show with extensive experimental results that it can be applied to density estimation on synthetic data, image generation, image denoising, and training latent variable models with implicit encoders.
**********

Debiasing Distributed Second Order Optimization with Surrogate Sketching and Scaled Regularization
Michal Derezinski, Burak Bartan, Mert Pilanci, Michael W. Mahoney
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Controlled Differential Equations for Irregular Time Series
Patrick Kidger, James Morrill, James Foster, Terry Lyons
Neural ordinary differential equations are an attractive option for modelling temporal dynamics. However, a fundamental issue is that the solution to an ordinary differential equation is determined by its initial condition, and there is no mechanism for adjusting the trajectory based on subsequent observations. Here, we demonstrate how this may be resolved through the well-understood mathematics of \emph{controlled differential equations}. The resulting \emph{neural controlled differential equation} model is directly applicable to the general setting of partially-observed irregularly-sampled multivariate time series, and (unlike previous work on this problem) it may utilise memory-efficient adjoint-based backpropagation even across observations. We demonstrate that our model achieves state-of-the-art performance against similar (ODE or RNN based) models in empirical studies on a range of datasets. Finally we provide theoretical results demonstrating universal approximation, and that our model subsumes alternative ODE models.
**********

On Efficiency in Hierarchical Reinforcement Learning
Zheng Wen, Doina Precup, Morteza Ibrahimi, Andre Barreto, Benjamin Van Roy, Satinder Singh
Hierarchical Reinforcement Learning (HRL) approaches promise to provide more efficient solutions to sequential decision making problems, both in terms of statistical as well as computational efficiency. While this has been demonstrated empirically over time in a variety of tasks, theoretical results quantifying the benefits of such methods are still few and far between. In this paper, we discuss the kind of structure in a Markov decision process which gives rise to efficient HRL methods. Specifically, we formalize the intuition that HRL can exploit well repeating "subMDPs", with similar reward and transition structure. We show that, under reasonable assumptions, a model-based Thompson sampling-style HRL algorithm that exploits this structure is statistically efficient, as established through a finite-time regret bound. We also establish conditions under which planning with structure-induced options is near-optimal and computationally efficient.
**********

On Correctness of Automatic Differentiation for Non-Differentiable Functions
Wonyeol Lee, Hangyeol Yu, Xavier Rival, Hongseok Yang
Differentiation lies at the core of many machine-learning algorithms, and is well-supported by popular autodiff systems, such as TensorFlow and PyTorch. Originally, these systems have been developed to compute derivatives of differentiable

functions, but in practice, they are commonly applied to functions with non-diff
erentiabilities. For instance, neural networks using ReLU define non-differentia
ble functions in general, but the gradients of losses involving those functions
are computed using autodiff systems in practice. This status quo raises a natura
l question: are autodiff systems correct in any formal sense when they are appli
ed to such non-differentiable functions? In this paper, we provide a positive an
swer to this question. Using counterexamples, we first point out flaws in often-
used informal arguments, such as: non-differentiabilities arising in deep learni
ng do not cause any issues because they form a measure-zero set. We then investi
gate a class of functions, called PAP functions, that includes nearly all (possi
bly non-differentiable) functions in deep learning nowadays. For these PAP funct
ions, we propose a new type of derivatives, called intensional derivatives, and
prove that these derivatives always exist and coincide with standard derivatives
 for almost all inputs. We also show that these intensional derivatives are what
 most autodiff systems compute or try to compute essentially. In this way, we fo
rmally establish the correctness of autodiff systems applied to non-differentiab
le functions.
**********

Probabilistic Linear Solvers for Machine Learning
Jonathan Wenger, Philipp Hennig
Linear systems are the bedrock of virtually all numerical computation. Machine l
earning poses specific challenges for the solution of such systems due to their
scale, characteristic structure, stochasticity and the central role of uncertain
ty in the field. Unifying earlier work we propose a class of probabilistic linea
r solvers which jointly infer the matrix, its inverse and the solution from matr
ix-vector product observations. This class emerges from a fundamental set of des
iderata which constrains the space of possible algorithms and recovers the metho
d of conjugate gradients under certain conditions. We demonstrate how to incorpo
rate prior spectral information in order to calibrate uncertainty and experiment
ally showcase the potential of such solvers for machine learning.
**********

Dynamic Regret of Policy Optimization in Non-Stationary Environments
Yingjie Fei, Zhuoran Yang, Zhaoran Wang, Qiaomin Xie
We consider reinforcement learning (RL) in episodic MDPs with adversarial full-i
nformation
  reward feedback and unknown fixed  transition kernels. We propose two model-fr
ee policy optimization algorithms,
  POWER and POWER++, and establish guarantees for their dynamic regret. Compared
  with the classical notion of static regret, dynamic regret is a stronger
  notion as it explicitly accounts for the non-stationarity of
  environments. The dynamic regret attained by the proposed
  algorithms interpolates between different regimes of non-stationarity,
  and moreover satisfies a notion of adaptive (near-)optimality, in
  the sense that it matches the (near-)optimal static regret under slow-changing
 environments.
  The dynamic regret bound features two components, one arising
  from exploration, which deals with the uncertainty of transition kernels,
  and the other arising from adaptation, which deals with non-stationary
  environments.
  Specifically, we show that POWER++ improves over POWER on the
  second component of the dynamic regret by  actively adapting
  to non-stationarity through prediction.
  To the best of our knowledge,
  our work is the first dynamic regret analysis of model-free RL algorithms
  in non-stationary environments.
**********

Multipole Graph Neural Operator for Parametric Partial Differential Equations
Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Andrew Stuart,
 Kaushik Bhattacharya, Anima Anandkumar
One of the main challenges in using deep learning-based methods for simulating p

hysical systems and solving partial differential equations (PDEs) is formulating physics-based data in the desired structure for neural networks. Graph neural networks (GNNs) have gained popularity in this area since graphs offer a natural way of modeling particle interactions and provide a clear way of discretizing the continuum models. However, the graphs constructed for approximating such tasks usually ignore long-range interactions due to unfavorable scaling of the computational complexity with respect to the number of nodes. The errors due to these approximations scale with the discretization of the system, thereby not allowing for generalization under mesh-refinement. Inspired by the classical multipole methods, we purpose a novel multi-level graph neural network framework that captures interaction at all ranges with only linear complexity. Our multi-level formulation is equivalent to recursively adding inducing points to the kernel matrix, unifying GNNs with multi-resolution matrix factorization of the kernel. Experiments confirm our multi-graph network learns discretization-invariant solution operators to PDEs and can be evaluated in linear time.

**********

BlockGAN: Learning 3D Object-aware Scene Representations from Unlabelled Images

Thu H. Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, Niloy Mitra

We present BlockGAN, an image generative model that learns object-aware 3D scene representations directly from unlabelled 2D images. Current work on scene representation learning either ignores scene background or treats the whole scene as one object. Meanwhile, work that considers scene compositionality treats scene objects only as image patches or 2D layers with alpha maps. Inspired by the computer graphics pipeline, we design BlockGAN to learn to first generate 3D features of background and foreground objects, then combine them into 3D features for the whole scene, and finally render them into realistic images. This allows BlockGAN to reason over occlusion and interaction between objects' appearance, such as shadow and lighting, and provides control over each object's 3D pose and identity, while maintaining image realism. BlockGAN is trained end-to-end, using only unlabelled single images, without the need for 3D geometry, pose labels, object masks, or multiple views of the same scene. Our experiments show that using explicit 3D features to represent objects allows BlockGAN to learn disentangled representations both in terms of objects (foreground and background) and their properties (pose and identity).

**********

Online Structured Meta-learning

Huaxiu Yao, Yingbo Zhou, Mehrdad Mahdavi, Zhenhui (Jessie) Li, Richard Socher, Caiming Xiong

Learning quickly is of great importance for machine intelligence deployed in online platforms. With the capability of transferring knowledge from learned tasks, meta-learning has shown its effectiveness in online scenarios by continuously updating the model with the learned prior. However, current online meta-learning algorithms are limited to learn a globally-shared meta-learner, which may lead to sub-optimal results when the tasks contain heterogeneous information that are difficult to share. We overcome this limitation by proposing an online structured meta-learning (OSML) framework. Inspired by the knowledge organization of human and hierarchical feature representation, OSML explicitly disentangles the meta-learner as a meta-hierarchical graph with different knowledge blocks. When a new task is encountered, it constructs a meta-knowledge pathway by either utilizing the most relevant knowledge blocks or exploring new blocks. Through the meta-knowledge pathway, the model is able to quickly adapt to the new task. In addition, new knowledge is further incorporated into the selected blocks. Experiments on three datasets empirically demonstrate the effectiveness and interpretability of our proposed framework, not only under heterogeneous tasks but also under homogeneous settings.

**********

Learning Strategic Network Emergence Games

Rakshit Trivedi, Hongyuan Zha

Real-world networks, especially the ones that emerge due to actions of animate agents (e.g. humans, animals), are the result of underlying strategic mechanisms

aimed at maximizing individual or collective benefits. Learning approaches built to capture these strategic insights would gain interpretability and flexibility benefits that are required to generalize beyond observations.
To this end, we consider a game-theoretic formalism of network emergence that accounts for the underlying strategic mechanisms and take it to the observed data.

We propose MINE (Multi-agent Inverse models of Network Emergence mechanism), a new learning framework that solves Markov-Perfect network emergence games using multi-agent inverse reinforcement learning. MINE jointly discovers agents' strategy profiles in the form of network emergence policy and the latent payoff mechanism in the form of learned reward function. In the experiments, we demonstrate that MINE learns versatile payoff mechanisms that: highly correlates with the ground truth for a synthetic case; can be used to analyze the observed network structure; and enable effective transfer in specific settings. Further, we show that the network emergence game as a learned model supports meaningful strategic predictions, thereby signifying its applicability to a variety of network analysis tasks.
**********
Towards Interpretable Natural Language Understanding with Explanations as Latent Variables
Wangchunshu Zhou, Jinyi Hu, Hanlin Zhang, Xiaodan Liang, Maosong Sun, Chenyan Xiong, Jian Tang
Recently generating natural language explanations has shown very promising results in not only offering interpretable explanations but also providing additional information and supervision for prediction. However, existing approaches usually require a large set of human annotated explanations for training while collecting a large set of explanations is not only time consuming but also expensive. In this paper, we develop a general framework for interpretable natural language understanding that requires only a small set of human annotated explanations for training. Our framework treats natural language explanations as latent variables that model the underlying reasoning process of a neural model. We develop a variational EM framework for optimization where an explanation generation module and an explanation-augmented prediction module are alternatively optimized and mutually enhance each other. Moreover, we further propose an explanation-based self-training method under this framework for semi-supervised learning. It alternates between assigning pseudo-labels to unlabeled data and generating new explanations to iteratively improve each other. Experiments on two natural language understanding tasks demonstrate that our framework can not only make effective predictions in both supervised and semi-supervised settings, but is also able to generate good natural language explanations.
**********
The Mean-Squared Error of Double Q-Learning
Wentao Weng, Harsh Gupta, Niao He, Lei Ying, R. Srikant
In this paper, we establish a theoretical comparison between the asymptotic mean square errors of double Q-learning and Q-learning. Our result builds upon an analysis for linear stochastic approximation based on Lyapunov equations and applies to both tabular setting or with linear function approximation, provided that the optimal policy is unique and the algorithms converge. We show that the asymptotic mean-square error of Double Q-learning is exactly equal to that of Q-learning if Double Q-learning uses twice the learning rate of Q-learning and the output of Double Q-learning is the average of its two estimators. We also present some practical implications of this theoretical observation using simulations.
**********
What Makes for Good Views for Contrastive Learning?
Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, Phillip Isola
Contrastive learning between multiple views of the data has recently achieved state of the art performance in the field of self-supervised representation learning. Despite its success, the influence of different view choices has been less studied. In this paper, we use theoretical and empirical analysis to better under

stand the importance of view selection, and argue that we should reduce the mutu
al information (MI) between views while keeping task-relevant information intact
.
To verify this hypothesis, we devise unsupervised and semi-supervised frameworks
 that learn effective views by aiming to reduce their MI. We also consider data
augmentation as a way to reduce MI, and show that increasing data augmentation i
ndeed leads to decreasing MI and improves downstream classification accuracy. As
 a by-product, we achieve a new state-of-the-art accuracy on unsupervised pre-tr
aining for ImageNet classification (73% top-1 linear readout with a ResNet-50).
**********

Denoising Diffusion Probabilistic Models
Jonathan Ho, Ajay Jain, Pieter Abbeel
We present high quality image synthesis results using diffusion probabilistic mo
dels, a class of latent variable models inspired by considerations from nonequil
ibrium thermodynamics. Our best results are obtained by training on a weighted v
ariational bound designed according to a novel connection between diffusion prob
abilistic models and denoising score matching with Langevin dynamics, and our mo
dels naturally admit a progressive lossy decompression scheme that can be interp
reted as a generalization of autoregressive decoding. On the unconditional CIFAR
10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID scor
e of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN.
**********

Barking up the right tree: an approach to search over molecule synthesis DAGs
John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin Segler, José Miguel Hernánde
z-Lobato
When designing new molecules with particular properties, it is not only importan
t what to make but crucially how to make it. These instructions form a synthesis
 directed acyclic graph (DAG), describing how a large vocabulary of simple build
ing blocks can be recursively combined through chemical reactions to create more
 complicated molecules of interest. In contrast, many current deep generative mo
dels for molecules ignore synthesizability. We therefore propose a deep generati
ve model that better represents the real world process, by directly outputting m
olecule synthesis DAGs. We argue that this provides sensible inductive biases, e
nsuring that our model searches over the same chemical space that chemists would
 also have access to, as well as interoperability. We show that our approach is
able to model chemical space well, producing a wide range of diverse molecules,
and allows for unconstrained optimization of an inherently constrained problem:
maximize certain chemical properties such that discovered molecules are synthesi
zable.
**********

On Uniform Convergence and Low-Norm Interpolation Learning
Lijia Zhou, Danica J. Sutherland, Nati Srebro
We consider an underdetermined noisy linear regression model where the minimum-n
orm interpolating predictor is known to be consistent,  and ask: can uniform con
vergence in a norm ball, or at least (following Nagarajan and Kolter) the subset
 of a norm ball that the algorithm selects on a typical input set, explain this
success? We show that uniformly bounding the difference between empirical and po
pulation errors cannot show any learning in the norm ball, and cannot show consi
stency for any set, even one depending on the exact algorithm and distribution.
But we argue we can explain the consistency of the minimal-norm interpolator wit
h a slightly weaker, yet standard, notion: uniform convergence of zero-error pre
dictors in a norm ball. We use this to bound the generalization error of low- (b
ut not minimal-)norm interpolating predictors.
**********

Bandit Samplers for Training Graph Neural Networks
Ziqi Liu, Zhengwei Wu, Zhiqiang Zhang, Jun Zhou, Shuang Yang, Le Song, Yuan Qi
Several sampling algorithms with variance reduction have been
proposed for accelerating the training of Graph Convolution Networks (GCNs). How
ever, due to the intractable computation of optimal sampling distribution,
these sampling algorithms are suboptimal for GCNs and are not

applicable to more general graph neural networks (GNNs) where
the message aggregator contains learned weights rather than
fixed weights, such as Graph Attention Networks (GAT).
The fundamental reason is that the embeddings of the neighbors or learned weight
s involved in the optimal sampling distribution
are \emph{changing} during the training and \emph{not known a priori},
but only \emph{partially observed} when sampled, thus
making the derivation of an optimal variance reduced samplers non-trivial.
In this paper, we formulate the optimization of the sampling
variance as an adversary bandit problem, where the rewards are related to the no
de embeddings and learned weights,
and can vary constantly. Thus a good sampler needs to acquire
variance information about more neighbors (exploration) while at the same time o
ptimizing the immediate sampling variance (exploit). We theoretically
show that our algorithm asymptotically approaches the optimal variance within a
factor of 3. We show the efficiency and effectiveness of our approach on multipl
e
datasets.
**********
Sampling from a k-DPP without looking at all items
Daniele Calandriello, Michal Derezinski, Michal Valko
Determinantal point processes (DPPs) are a useful probabilistic model for select
ing
a small diverse subset out of a large collection of items, with applications in
summarization, recommendation, stochastic optimization, experimental design and
more. Given a kernel function and a subset size k, our goal is to sample k out
of n items with probability proportional to the determinant of the kernel matrix
induced by the subset (a.k.a. k-DPP). Existing k-DPP sampling algorithms require
an expensive preprocessing step which involves multiple passes over all n items,
making it infeasible for large datasets. A naïve heuristic addressing this probl
em is
to uniformly subsample a fraction of the data and perform k-DPP sampling only on
those items, however this method offers no guarantee that the produced sample wi
ll
even approximately resemble the target distribution over the original dataset. I
n this
paper, we develop alpha-DPP, an algorithm which adaptively builds a sufficiently
 large uniform
sample of data that is then used to efficiently generate a smaller set of k item
s,
while ensuring that this set is drawn exactly from the target distribution defin
ed
on all n items. We show empirically that our
algorithm produces a k-DPP sample after observing only a small fraction of all
elements, leading to several orders of magnitude faster performance compared to
the state-of-the-art. Our implementation of alpha-DPP is provided at https://git
hub.com/guilgautier/DPPy/.
**********
Uncovering the Topology of Time-Varying fMRI Data using Cubical Persistence
Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicho
las Turk-Browne, Smita Krishnaswamy
Functional magnetic resonance imaging (fMRI) is a crucial technology for gaining
 insights into cognitive processes in humans. Data amassed from fMRI measurement
s result in volumetric data sets that vary over time. However, analysing such da
ta presents a challenge due to the large degree of noise and person-to-person va
riation in how information is represented in the brain. To address this challeng
e, we present a novel topological approach that encodes each time point in an fM
RI data set as a persistence diagram of topological features, i.e. high-dimensio
nal voids present in the data. This representation naturally does not rely on vo
xel-by-voxel correspondence and is robust towards noise. We show that these time

-varying persistence diagrams can be clustered to find meaningful groupings betw
een participants, and that they are also useful in studying within-subject brain
 state trajectories of subjects performing a particular task. Here, we apply bot
h clustering and trajectory analysis techniques to a group of participants watch
ing the movie 'Partly Cloudy'. We observe significant differences in both brain
state trajectories and overall topological activity between adults and children
watching the same movie.
**********

Hierarchical Poset Decoding for Compositional Generalization in Language
Yinuo Guo, Zeqi Lin, Jian-Guang Lou, Dongmei Zhang
We formalize human language understanding as a structured prediction task where
the output is a partially ordered set (poset). Current encoder-decoder architect
ures do not take the poset structure of semantics into account properly, thus su
ffering from poor compositional generalization ability. In this paper, we propos
e a novel hierarchical poset decoding paradigm for compositional generalization
in language. Intuitively: (1) the proposed paradigm enforces partial permutation
 invariance in semantics, thus avoiding overfitting to bias ordering information
; (2) the hierarchical mechanism allows to capture high-level structures of pose
ts. We evaluate our proposed decoder on Compositional Freebase Questions (CFQ),
a large and realistic natural language question answering dataset that is specif
ically designed to measure compositional generalization. Results show that it ou
tperforms current decoders.
**********

Evaluating and Rewarding Teamwork Using Cooperative Game Abstractions
Tom Yan, Christian Kroer, Alexander Peysakhovich
Can we predict how well a team of individuals will perform together? How should
individuals be rewarded for their  contributions to the team performance? Cooper
ative game theory gives us a powerful set of tools for answering these questions
: the characteristic function and solution concepts like the Shapley Value. Ther
e are two major difficulties in applying these techniques to real world problems
: first, the characteristic function is rarely given to us and needs to be learn
ed from data. Second, the Shapley Value is combinatorial in nature. We introduce
 a parametric model called cooperative game abstractions (CGAs) for estimating c
haracteristic functions from data. CGAs are easy to learn, readily interpretable
, and crucially allows linear-time computation of the Shapley Value. We provide
identification results and sample complexity bounds for CGA models as well as er
ror bounds in the estimation of the Shapley Value using CGAs. We apply our metho
ds to study teams of artificial RL agents as well as real world teams from profe
ssional sports.
**********

Exchangeable Neural ODE for Set Modeling
Yang Li, Haidong Yi, Christopher Bender, Siyuan Shan, Junier B. Oliva
Reasoning over an instance composed of a set of vectors, like a point cloud, req
uires that one accounts for intra-set dependent features among elements. However
, since such instances are unordered, the elements' features should remain uncha
nged when the input's order is permuted. This property, permutation equivariance
, is a challenging constraint for most neural architectures. While recent work h
as proposed global pooling and attention-based solutions, these may be limited i
n the way that intradependencies are captured in practice. In this work we propo
se a more general formulation to achieve permutation equivariance through ordina
ry differential equations (ODE). Our proposed module, Exchangeable Neural ODE (E
xNODE), can be seamlessly applied for both discriminative and generative tasks.
We also extend set modeling in the temporal dimension and propose a VAE based mo
del for temporal set modeling. Extensive experiments demonstrate the efficacy of
 our method over strong baselines.
**********

Profile Entropy: A Fundamental Measure for the Learnability and Compressibility
of Distributions
Yi Hao, Alon Orlitsky
The profile of a sample is the multiset of its symbol frequencies. We show that

for samples of discrete distributions, profile entropy is a fundamental measure unifying the concepts of estimation, inference, and compression. Specifically, profile entropy: a) determines the speed of estimating the distribution relative to the best natural estimator; b) characterizes the rate of inferring all symmetric properties compared with the best estimator over any label-invariant distribution collection; c) serves as the limit of profile compression, for which we derive optimal near-linear-time block and sequential algorithms. To further our understanding of profile entropy, we investigate its attributes, provide algorithms for approximating its value, and determine its magnitude for numerous structural distribution families.
**********

CoADNet: Collaborative Aggregation-and-Distribution Networks for Co-Salient Object Detection
Qijian Zhang, Runmin Cong, Junhui Hou, Chongyi Li, Yao Zhao
Co-Salient Object Detection (CoSOD) aims at discovering salient objects that repeatedly appear in a given query group containing two or more relevant images. One challenging issue is how to effectively capture co-saliency cues by modeling and exploiting inter-image relationships. In this paper, we present an end-to-end collaborative aggregation-and-distribution network (CoADNet) to capture both salient and repetitive visual patterns from multiple images. First, we integrate saliency priors into the backbone features to suppress the redundant background information through an online intra-saliency guidance structure. After that, we design a two-stage aggregate-and-distribute architecture to explore group-wise semantic interactions and produce the co-saliency features. In the first stage, we propose a group-attentional semantic aggregation module that models inter-image relationships to generate the group-wise semantic representations. In the second stage, we propose a gated group distribution module that adaptively distributes the learned group semantics to different individuals in a dynamic gating mechanism. Finally, we develop a group consistency preserving decoder tailored for the CoSOD task, which maintains group constraints during feature decoding to predict more consistent full-resolution co-saliency maps. The proposed CoADNet is evaluated on four prevailing CoSOD benchmark datasets, which demonstrates the remarkable performance improvement over ten state-of-the-art competitors.
**********

Regularized linear autoencoders recover the principal components, eventually
Xuchan Bao, James Lucas, Sushant Sachdeva, Roger B. Grosse
Our understanding of learning input-output relationships with neural nets has improved rapidly in recent years, but little is known about the convergence of the underlying representations, even in the simple case of linear autoencoders (LAEs). We show that when trained with proper regularization, LAEs can directly learn the optimal representation -- ordered, axis-aligned principal components. We analyze two such regularization schemes: non-uniform L2 regularization and a deterministic variant of nested dropout [Rippel et al, ICML' 2014]. Though both regularization schemes converge to the optimal representation, we show that this convergence is slow due to ill-conditioning that worsens with increasing latent dimension. We show that the inefficiency of learning the optimal representation is not inevitable -- we present a simple modification to the gradient descent update that greatly speeds up convergence empirically.
**********

Semi-Supervised Partial Label Learning via Confidence-Rated Margin Maximization
Wei Wang, Min-Ling Zhang
Partial label learning assumes inaccurate supervision where each training example is associated with a set of candidate labels, among which only one is valid. In many real-world scenarios, however, it is costly and time-consuming to assign candidate label sets to all the training examples. To circumvent this difficulty, the problem of semi-supervised partial label learning is investigated in this paper, where unlabeled data is utilized to facilitate model induction along with partial label training examples. Specifically, label propagation is adopted to instantiate the labeling confidence of partial label examples. After that, maximum margin formulation is introduced to jointly enable the induction of predictiv

e model and the estimation of labeling confidence over unlabeled data. The deriv ed formulation enforces confidence-rated margin maximization and confidence mani fold preservation over partial label examples and unlabeled data. We show that t he predictive model and labeling confidence can be solved via alternating optimi zation which admits QP solutions in either alternating step. Extensive experimen ts on synthetic as well as real-world data sets clearly validate the effectiven ess of the proposed semi-supervised partial label learning approach.
**********

GramGAN: Deep 3D Texture Synthesis From 2D Exemplars
Tiziano Portenier, Siavash Arjomand Bigdeli, Orcun Goksel
We present a novel texture synthesis framework, enabling the generation of infin ite, high-quality 3D textures given a 2D exemplar image. Inspired by recent adva nces in natural texture synthesis, we train deep neural models to generate textu res by non-linearly combining learned noise frequencies. To achieve a highly rea listic output conditioned on an exemplar patch, we propose a novel loss function that combines ideas from both style transfer and generative adversarial network s. In particular, we train the synthesis network to match the Gram matrices of d eep features from a discriminator network. In addition, we propose two architect ural concepts and an extrapolation strategy that significantly improve generaliz ation performance. In particular, we inject both model input and condition into hidden network layers by learning to scale and bias hidden activations. Quantita tive and qualitative evaluations on a diverse set of exemplars motivate our desi gn decisions and show that our system performs superior to previous state of the art. Finally, we conduct a user study that confirms the benefits of our framewo rk.
**********

UWSOD: Toward Fully-Supervised-Level Capacity Weakly Supervised Object Detection
Yunhang Shen, Rongrong Ji, Zhiwei Chen, Yongjian Wu, Feiyue Huang
Weakly supervised object detection (WSOD) has attracted extensive research atten tion due to its great flexibility of exploiting large-scale dataset with only im age-level annotations for detector training. Despite its great advance in recent years, WSOD still suffers limited performance, which is far below that of fully supervised object detection (FSOD). As most WSOD methods depend on object propo sal algorithms to generate candidate regions and are also confronted with challe nges like low-quality predicted bounding boxes and large scale variation. In thi s paper, we propose a unified WSOD framework, termed UWSOD, to develop a high-ca pacity general detection model with only image-level labels, which is self-conta ined and does not require external modules or additional supervision. To this en d, we exploit three important components, i.e., object proposal generation, boun ding-box fine-tuning and scale-invariant features. First, we propose an anchor-b ased self-supervised proposal generator to hypothesize object locations, which i s trained end-to-end with supervision created by UWSOD for both objectness class ification and regression. Second, we develop a step-wise bounding-box fine-tunin g to refine both detection scores and coordinates by progressively select high-c onfidence object proposals as positive samples, which bootstraps the quality of predicted bounding boxes. Third, we construct a multi-rate resampling pyramid to aggregate multi-scale contextual information, which is the first in-network fea ture hierarchy to handle scale variation in WSOD. Extensive experiments on PASCA L VOC and MS COCO show that the proposed UWSOD achieves competitive results with the state-of-the-art WSOD methods while not requiring external modules or addit ional supervision. Moreover, the upper-bound performance of UWSOD with class-agn ostic ground-truth bounding boxes approaches Faster R-CNN, which demonstrates UW SOD has fully-supervised-level capacity.
**********

Learning Restricted Boltzmann Machines with Sparse Latent Variables
Guy Bresler, Rares-Darius Buhai
Requests for name changes in the electronic proceedings will be accepted with 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 Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction
Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, 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-authors prior to requesting a name change in the electronic proceedings.
**********
Curriculum learning for multilevel budgeted combinatorial problems
Adel Nabli, Margarida Carvalho
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
FedSplit: an algorithmic framework for fast federated optimization
Reese Pathak, Martin J. Wainwright
Motivated by federated learning, we consider the hub-and-spoke model of distributed optimization in which a central authority coordinates the computation of a solution among many agents while limiting communication. We first study some past procedures for federated optimization, and show that their fixed points need not correspond to stationary points of the original optimization problem, even in simple convex settings with deterministic updates. In order to remedy these issues, we introduce FedSplit, a class of algorithms based on operator splitting procedures for solving distributed convex minimization with additive structure. We prove that these procedures have the correct fixed points, corresponding to optima of the original optimization problem, and we characterize their convergence rates under different settings. Our theory shows that these methods are provably robust to inexact computation of intermediate local quantities. We complement our theory with some experiments that demonstrate the benefits of our methods in practice.
**********
Estimation and Imputation in Probabilistic Principal Component Analysis with Missing Not At Random Data
Aude Sportisse, Claire Boyer, Julie Josse
Missing Not At Random (MNAR) values where the probability of having missing data may depend on the missing value itself, are notoriously difficult to account for in analyses, although very frequent in the data. One solution to handle MNAR data is to specify a model for the missing data mechanism, which makes inference or imputation tasks more complex.
Furthermore, this implies a strong \textit{a priori} on the parametric form of the distribution.
However, some works have obtained guarantees on the estimation of parameters in the presence of MNAR data, without specifying the distribution of missing data \citep{mohan2018estimation, tang2003analysis}. This is very useful in practice, but is limited to simple cases such as few self-masked MNAR variables in data generated according to linear regression models.
We continue this line of research, but extend it to a more general MNAR mechanism, in a more general model of the probabilistic principal component analysis (PPCA), \textit{i.e.}, a low-rank model with random effects. We prove identifiability of the PPCA parameters. We then propose an estimation of the loading coefficients and a data imputation method. They are based on estimators of means, variances and covariances of missing variables, for which consistency is discussed. These estimators have the great advantage of being calculated using only the observed information, leveraging the underlying low-rank structure of the data. We illustrate the relevance of the method with numerical experiments on synthetic data and also on two datasets, one collected from a medical register and the other one from a recommendation system.
**********

Correlation Robust Influence Maximization
Louis Chen, Divya Padmanabhan, Chee Chin Lim, Karthik Natarajan
We propose a distributionally robust model for the influence maximization proble
m. Unlike the classical independent cascade model of Kempe et al (2003), this mo
del's diffusion process is adversarially adapted to the choice of seed set. So i
nstead of optimizing under the assumption that all influence relationships in th
e network are independent, we seek a seed set whose expected influence under the
 worst correlation, i.e., the ``worst-case, expected influence", is maximized. W
e show that this worst-case influence can be efficiently computed, and though th
e optimization is NP-hard, a (1 - 1/e) approximation guarantee holds. We also an
alyze the structure to the adversary's choice of diffusion process, and contras
t with established models. Beyond the key computational advantages, we also stud
y the degree to which the independence assumption may be considered costly, and
provide insights from numerical experiments comparing the adversarial and indepe
ndent cascade model.
**********

Neuronal Gaussian Process Regression
Johannes Friedrich
The brain takes uncertainty intrinsic to our world into account. For example, as
sociating spatial locations with rewards requires to predict not only expected r
eward at new spatial locations but also its uncertainty to avoid catastrophic ev
ents and forage safely. A powerful and flexible framework for nonlinear regressi
on that takes uncertainty into account in a principled Bayesian manner is Gaussi
an process (GP) regression. Here I propose that the brain implements GP regressi
on and present neural networks (NNs) for it. First layer neurons, e.g.\ hippocam
pal place cells, have tuning curves that correspond to evaluations of the GP ker
nel. Output neurons explicitly and distinctively encode predictive mean and vari
ance, as observed in orbitofrontal cortex (OFC) for the case of reward predictio
n. Because the weights of a NN implementing exact GP regression do not arise wit
h biological plasticity rules, I present approximations to obtain local (anti-)H
ebbian synaptic learning rules. The resulting neuronal network approximates the
full GP well compared to popular sparse GP approximations and achieves comparabl
e predictive performance.
**********

Nonconvex Sparse Graph Learning under Laplacian Constrained Graphical Model
Jiaxi Ying, José Vinícius de Miranda Cardoso , Daniel Palomar
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Synthetic Data Generators -- Sequential and Private
Olivier Bousquet, Roi Livni, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Uncertainty Quantification for Inferring Hawkes Networks
Haoyun Wang, Liyan Xie, Alex Cuozzo, Simon Mak, Yao Xie
Multivariate Hawkes processes are commonly used to model streaming networked eve
nt data in a wide variety of applications. However, it remains a challenge to ex
tract reliable inference from complex datasets with uncertainty quantification.
Aiming towards this, we develop a statistical inference framework to learn causa
l relationships between nodes from networked data, where the underlying directed
 graph implies Granger causality. We provide uncertainty quantification for the
maximum likelihood estimate of the network multivariate Hawkes process by provid
ing a non-asymptotic confidence set. The main technique is based on the concentr
ation inequalities of continuous-time martingales. We compare our method to the
previously-derived asymptotic Hawkes process confidence interval, and demonstrat

e the strengths of our method in an application to neuronal connectivity reconst
ruction.
**********

Implicit Distributional Reinforcement Learning
Yuguang Yue, Zhendong Wang, Mingyuan Zhou
To improve the sample efficiency of policy-gradient based reinforcement learning
 algorithms, we propose implicit distributional actor-critic (IDAC) that consist
s of a distributional critic, built on two deep generator networks (DGNs), and a
 semi-implicit actor (SIA), powered by a flexible policy distribution. We adopt
a distributional perspective on the discounted cumulative return and model it wi
th a state-action-dependent implicit distribution, which is approximated by the
DGNs that take state-action pairs and random noises as their input. Moreover, we
 use the SIA to provide a semi-implicit policy distribution, which mixes the pol
icy parameters with a reparameterizable distribution that is not constrained by
an analytic density function. In this way, the policy's marginal distribution is
 implicit, providing the potential to model complex properties such as covarianc
e structure and skewness, but its parameter and entropy can still be estimated.
We incorporate these features with an off-policy algorithm framework to solve pr
oblems with continuous action space and compare IDAC with state-of-the-art algor
ithms on representative OpenAI Gym environments. We observe that IDAC outperform
s these baselines in most tasks. Python code is provided.
**********

Auxiliary Task Reweighting for Minimum-data Learning
Baifeng Shi, Judy Hoffman, Kate Saenko, Trevor Darrell, Huijuan Xu
Supervised learning requires a large amount of training data, limiting its appli
cation where labeled data is scarce. To compensate for data scarcity, one possib
le method is to utilize auxiliary tasks to provide additional supervision for th
e main task. Assigning and optimizing the importance weights for different auxil
iary tasks remains an crucial and largely understudied research question. In thi
s work, we propose a method to automatically reweight auxiliary tasks in order t
o reduce the data requirement on the main task. Specifically, we formulate the w
eighted likelihood function of auxiliary tasks as a surrogate prior for the main
 task. By adjusting the auxiliary task weights to minimize the divergence betwee
n the surrogate prior and the true prior of the main task, we obtain a more accu
rate prior estimation, achieving the goal of minimizing the required amount of t
raining data for the main task and avoiding a costly grid search. In multiple ex
perimental settings (e.g. semi-supervised learning, multi-label classification),
 we demonstrate that our algorithm can effectively utilize limited labeled data
of the main task with the benefit of auxiliary tasks compared with previous task
 reweighting methods. We also show that under extreme cases with only a few extr
a examples (e.g. few-shot domain adaptation), our algorithm results in significa
nt improvement over the baseline. Our code and video is available at https://sit
es.google.com/view/auxiliary-task-reweighting.
**********

Small Nash Equilibrium Certificates in Very Large Games
Brian Zhang, Tuomas Sandholm
In many game settings, the game is not explicitly given but is only accessible b
y playing it. While there have  been impressive demonstrations in such settings,
 prior techniques have not offered safety guarantees, that is, guarantees on the
 game-theoretic exploitability of the computed strategies. In this paper we intr
oduce an approach that shows that it is possible to provide exploitability guara
ntees in such settings without ever exploring the entire game. We introduce a no
tion of a certificatae of an extensive-form approximate Nash equilibrium. For ve
rifying a certificate, we give an algorithm that runs in time linear in the size
 of the certificate rather than the size of the whole game. In zero-sum games, w
e further show that an optimal certificate---given the exploration so far---can
be computed with any standard game-solving algorithm (e.g., using a linear progr
am or counterfactual regret minimization). However, unlike in the cases of norma
l form or perfect information, we show that certain families of extensive-form g
ames do not have small approximate certificates, even after making extremely nic

e assumptions on the structure of the game. Despite this difficulty, we find exp
erimentally that very small certificates, even exact ones, often exist in large
and even in infinite games. Overall, our approach enables one to try one's favor
ite exploration strategies while offering exploitability guarantees, thereby dec
oupling the exploration strategy from the equilibrium-finding process.
**********

Training Linear Finite-State Machines
Arash Ardakani, Amir Ardakani, Warren Gross
A finite-state machine (FSM) is a computation model to process binary strings in
 sequential circuits. Hence, a single-input linear FSM is conventionally used to
 implement complex single-input functions , such as tanh and exponentiation func
tions, in stochastic computing (SC) domain where continuous values are represent
ed by sequences of random bits. In this paper, we introduce a method that can tr
ain a multi-layer FSM-based network where FSMs are connected to every FSM in the
 previous and the next layer. We show that the proposed FSM-based network can sy
nthesize multi-input complex functions such as 2D Gabor filters and can perform
non-sequential tasks such as image classifications on stochastic streams with no
 multiplication since FSMs are implemented by look-up tables only. Inspired by t
he capability of FSMs in processing binary streams, we then propose an FSM-based
 model that can process time series data when performing temporal tasks such as
character-level language modeling. Unlike long short-term memories (LSTMs) that
unroll the network for each input time step and perform back-propagation on the
unrolled network, our FSM-based model requires to backpropagate gradients only f
or the current input time step while it is still capable of learning long-term d
ependencies. Therefore, our FSM-based model can learn extremely long-term depend
encies as it requires 1/l memory storage during training compared to LSTMs, wher
e l is the number of time steps. Moreover, our FSM-based model reduces the power
 consumption of training on a GPU by 33% compared to an LSTM model of the same s
ize.
**********

Efficient active learning of sparse halfspaces with arbitrary bounded noise
Chicheng Zhang, Jie Shen, Pranjal Awasthi
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Swapping Autoencoder for Deep Image Manipulation
Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros,
 Richard Zhang
Deep generative models have become increasingly effective at producing realistic
 images from randomly sampled seeds, but using such models for controllable mani
pulation of existing images remains challenging. We propose the Swapping Autoenc
oder, a deep model designed specifically for image manipulation, rather than ran
dom sampling. The key idea is to encode an image into two independent components
 and enforce that any swapped combination maps to a realistic image. In particul
ar, we encourage the components to represent structure and texture, by enforcing
 one component to encode co-occurrent patch statistics across different parts of
 the image. As our method is trained with an encoder, finding the latent codes f
or a new input image becomes trivial, rather than cumbersome. As a result,  our
method enables us to manipulate real input images in various ways, including tex
ture swapping, local and global editing, and latent code vector arithmetic. Expe
riments on multiple datasets show that our model produces better results and is
substantially more efficient compared to recent generative models.
**********

Self-Supervised Few-Shot Learning on Point Clouds
Charu Sharma, Manohar Kaul
The increased availability of massive point clouds coupled with their utility in
 a wide variety of applications such as robotics, shape synthesis, and self-driv
ing cars has attracted increased attention from both industry and academia. Rece

ntly, deep neural networks operating on labeled point clouds have shown promisin
g results on supervised learning tasks like classification and segmentation. How
ever, supervised learning leads to the cumbersome task of annotating the point c
louds. To combat this problem, we propose two novel self-supervised pre-training
 tasks that encode a hierarchical partitioning of the point clouds using a cover
-tree, where point cloud subsets lie within balls of varying radii at each level
 of the cover-tree. Furthermore, our self-supervised learning network is restric
ted to pre-train on the support set (comprising of scarce training examples) use
d to train the downstream network in a few-shot learning (FSL) setting. Finally,
 the fully-trained self-supervised network's point embeddings are input to the d
ownstream task's network. We present a comprehensive empirical evaluation of our
 method on both downstream classification and segmentation tasks and show that s
upervised methods pre-trained with our self-supervised learning method significa
ntly improve the accuracy of state-of-the-art methods. Additionally, our method
also outperforms previous unsupervised methods in downstream classification task
s.
**********

Faster Differentially Private Samplers via Rényi Divergence Analysis of Discreti
zed Langevin MCMC
Arun Ganesh, Kunal Talwar
Requests for name changes in the electronic proceedings will be accepted with 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 identifiable and interpretable latent models of high-dimensional neural
 activity using pi-VAE
Ding Zhou, Xue-Xin Wei
The ability to record activities from hundreds of neurons simultaneously in the
brain has placed an increasing demand for developing appropriate statistical tec
hniques to analyze such data. Recently, deep generative models have been propose
d to fit neural population responses. While these methods are flexible and expre
ssive, the downside is that they can be difficult to interpret and identify. To
address this problem, we propose a method that integrates key ingredients from l
atent models and traditional neural encoding models. Our method, pi-VAE, is insp
ired by recent progress on identifiable variational auto-encoder, which we adapt
 to make appropriate for neuroscience applications. Specifically, we propose to
construct latent variable models of neural activity while simultaneously modelin
g the relation between the latent and task variables (non-neural variables, e.g.
 sensory, motor, and other externally observable states). The incorporation of t
ask variables results in models that are not only more constrained, but also sho
w qualitative improvements in interpretability and identifiability. We validate
pi-VAE using synthetic data, and apply it to analyze neurophysiological datasets
 from rat hippocampus and macaque motor cortex. We demonstrate that pi-VAE not o
nly fits the data better, but also provides unexpected novel insights into the s
tructure of the neural codes.
**********

RL Unplugged: A Suite of Benchmarks for Offline Reinforcement Learning
Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Thomas Paine, Sergio Gómez, Konra
d Zolna, Rishabh Agarwal, Josh S. Merel, Daniel J. Mankowitz, Cosmin Paduraru, G
abriel Dulac-Arnold, Jerry Li, Mohammad Norouzi, Matthew Hoffman, Nicolas Heess,
 Nando de Freitas
Offline methods for reinforcement learning have a potential to help bridge the g
ap between reinforcement learning research and real-world applications. They mak
e it possible to learn policies from offline datasets, thus overcoming concerns
associated with online data collection in the real-world, including cost, safety
, or ethical concerns.  In this paper, we propose a benchmark called RL Unplugge
d to evaluate and compare offline RL methods. RL Unplugged includes data from a
diverse range of domains including games e.g., Atari benchmark) and simulated mo
tor control problems (e.g., DM Control Suite). The datasets include domains that

are partially or fully observable, use continuous or discrete actions, and hav
e stochastic vs. deterministic dynamics. We propose detailed evaluation protocol
s for each domain in RL Unplugged and provide an extensive analysis of supervise
d learning and offline RL methods using these protocols. We will release data fo
r all our tasks and open-source all algorithms presented in this paper. We hope
that our suite of benchmarks will increase the reproducibility of experiments an
d make it possible to study challenging tasks with a limited computational budge
t, thus making RL research both more systematic and more accessible across the c
ommunity. Moving forward, we view RL Unplugged as a living benchmark suite that
will evolve and grow with datasets contributed by the research community and our
selves. Our project page is available on github.
**********

Dual T: Reducing Estimation Error for Transition Matrix in Label-noise Learning
Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, Masashi S
ugiyama

Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Interior Point Solving for LP-based prediction+optimisation
Jayanta Mandi, Tias Guns

Solving optimization problem is the key to decision making in many real-life ana
lytics applications. However, the coefficients of the optimization problems are
often uncertain and dependent on external factors, such as future demand or ener
gy- or stock prices. Machine learning (ML) models, especially neural networks, a
re increasingly being used to estimate these coefficients in a data-driven way.
Hence, end-to-end predict-and-optimize approaches, which consider how effective
the predicted values are to solve the optimization problem, have received increa
sing attention. In case of integer linear programming problems, a popular approa
ch to overcome their non-differentiabilty is to add a quadratic penalty term to
the continuous relaxation, such that results from differentiating over quadratic
programs can be used. Instead we investigate the use of the more principled log
arithmic barrier term, as widely used in interior point solvers for linear progr
amming. Instead of differentiating the KKT conditions, we consider the homogeneo
us self-dual formulation of the LP and we show the relation between the interior
point step direction and corresponding gradients needed for learning. Finally,
our empirical experiments demonstrate our approach performs as good as if not be
tter than the state-of-the-art QPTL (Quadratic Programming task loss) formulatio
n of Wilder et al. and SPO approach of Elmachtoub and Grigas.
**********

A simple normative network approximates local non-Hebbian learning in the cortex
Siavash Golkar, David Lipshutz, Yanis Bahroun, Anirvan Sengupta, Dmitri Chklovsk
ii

To guide behavior, the brain extracts relevant features from high-dimensional da
ta streamed by sensory organs. Neuroscience experiments demonstrate that the pro
cessing of sensory inputs by cortical neurons is modulated by instructive signal
s which provide context and task-relevant information. Here, adopting a normativ
e approach, we model these instructive signals as supervisory inputs guiding the
projection of the feedforward data. Mathematically, we start with a family of R
educed-Rank Regression (RRR) objective functions which include Reduced Rank (min
imum) Mean Square Error (RRMSE) and Canonical Correlation Analysis (CCA), and de
rive novel offline and online optimization algorithms, which we call Bio-RRR. Th
e online algorithms can be implemented by neural networks whose synaptic learnin
g rules resemble calcium plateau potential dependent plasticity observed in the
cortex. We detail how, in our model, the calcium plateau potential can be inte
rpreted as a backpropagating error signal. We demonstrate that, despite relying
exclusively on biologically plausible local learning rules, our algorithms perfo
rm competitively with existing implementations of RRMSE and CCA.
**********

Kernelized information bottleneck leads to biologically plausible 3-factor Hebbian learning in deep networks
Roman Pogodin, Peter Latham
The state-of-the art machine learning approach to training deep neural networks, backpropagation, is implausible for real neural networks: neurons need to know their outgoing weights; training alternates between a bottom-up forward pass (computation) and a top-down backward pass (learning); and the algorithm often needs precise labels of many data points. Biologically plausible approximations to backpropagation, such as feedback alignment, solve the weight transport problem, but not the other two. Thus, fully biologically plausible learning rules have so far remained elusive. Here we present a family of learning rules that does not suffer from any of these problems. It is motivated by the information bottleneck principle (extended with kernel methods), in which networks learn to compress the input as much as possible without sacrificing prediction of the output. The resulting rules have a 3-factor Hebbian structure: they require pre- and post-synaptic firing rates and an error signal - the third factor - consisting of a global teaching signal and a layer-specific term, both available without a top-down pass. They do not require precise labels; instead, they rely on the similarity between pairs of desired outputs. Moreover, to obtain good performance on hard problems and retain biological plausibility, our rules need divisive normalization - a known feature of biological networks. Finally, simulations show that our rules perform nearly as well as backpropagation on image classification tasks.
**********
Understanding the Role of Training Regimes in Continual Learning
Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, Hassan Ghasemzadeh
Catastrophic forgetting affects the training of neural networks, limiting their ability to learn multiple tasks sequentially. From the perspective of the well established plasticity-stability dilemma, neural networks tend to be overly plastic, lacking the stability necessary to prevent the forgetting of previous knowledge, which means that as learning progresses, networks tend to forget previously seen tasks. This phenomenon coined in the continual learning literature, has attracted much attention lately, and several families of approaches have been proposed with different degrees of success. However, there has been limited prior work extensively analyzing the impact that different training regimes -- learning rate, batch size, regularization method-- can have on forgetting. In this work, we depart from the typical approach of altering the learning algorithm to improve stability. Instead, we hypothesize that the geometrical properties of the local minima found for each task play an important role in the overall degree of forgetting. In particular, we study the effect of dropout, learning rate decay, and batch size on forming training regimes that widen the tasks' local minima and consequently, on helping it not to forget catastrophically. Our study provides practical insights to improve stability via simple yet effective techniques that outperform alternative baselines.
**********
Fair regression with Wasserstein barycenters
Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, Massimiliano Pontil
We study the problem of learning a real-valued function that satisfies the Demographic Parity constraint. It demands the distribution of the predicted output to be independent of the sensitive attribute. We consider the case that the sensitive attribute is available for prediction. We establish a connection between fair regression and optimal transport theory, based on which we derive a close form expression for the optimal fair predictor.  Specifically, we show that the distribution of this optimum is the Wasserstein barycenter of the distributions induced by the standard regression function on the sensitive groups. This result offers an intuitive interpretation of the optimal fair prediction and suggests a simple post-processing algorithm to achieve fairness. We establish risk and distribution-free fairness guarantees for this procedure. Numerical experiments indicate that our method is very effective in learning fair models, with a relative increase in error rate that is inferior to the relative gain in fairness.

**********
Training Stronger Baselines for Learning to Optimize
Tianlong Chen, Weiyi Zhang, Zhou Jingyang, Shiyu Chang, Sijia Liu, Lisa Amini, Zhangyang Wang
Learning to optimize (L2O) is gaining increased attention because classical optimizers require laborious, problem-specific design and hyperparameter tuning. However, there are significant performance and practicality gaps between manually designed optimizers and existing L2O models. Specifically, learned optimizers are applicable to only a limited class of problems, often exhibit instability, and generalize poorly. As research efforts focus on increasingly sophisticated L2O models, we argue for an orthogonal, under-explored theme: improved training techniques for L2O models. We first present a progressive, curriculum-based training scheme, which gradually increases the optimizer unroll length to mitigate the well-known L2O dilemma of truncation bias (shorter unrolling) versus gradient explosion (longer unrolling). Secondly, we present an off-policy imitation learning based approach to guide the L2O learning, by learning from the behavior of analytical optimizers. We evaluate our improved training techniques with a variety of state-of-the-art L2O models and immediately boost their performance, without making any change to their model structures. We demonstrate that, using our improved training techniques, one of the earliest and simplest L2O models can be trained to outperform even the latest and most complex L2O models on a number of tasks. Our results demonstrate a greater potential of L2O yet to be unleashed, and prompt a reconsideration of recent L2O model progress. Our codes are publicly available at: https://github.com/VITA-Group/L2O-Training-Techniques.
**********
Exactly Computing the Local Lipschitz Constant of ReLU Networks
Matt Jordan, Alexandros G. Dimakis
The local Lipschitz constant of a neural network is a useful metric with applications in robustness, generalization, and fairness evaluation. We provide novel analytic results relating the local Lipschitz constant of nonsmooth vector-valued functions to a maximization over the norm of the generalized Jacobian. We present a sufficient condition for which backpropagation always returns an element of the generalized Jacobian, and reframe the problem over this broad class of functions. We show strong inapproximability results for estimating Lipschitz constants of ReLU networks, and then formulate an algorithm to compute these quantities exactly. We leverage this algorithm to evaluate the tightness of competing Lipschitz estimators and the effects of regularized training on the Lipschitz constant.
**********
Strictly Batch Imitation Learning by Energy-based Distribution Matching
Daniel Jarrett, Ioana Bica, Mihaela van der Schaar
Consider learning a policy purely on the basis of demonstrated behavior---that is, with no access to reinforcement signals, no knowledge of transition dynamics, and no further interaction with the environment. This strictly batch imitation learning problem arises wherever live experimentation is costly, such as in healthcare. One solution is simply to retrofit existing algorithms for apprenticeship learning to work in the offline setting. But such an approach leans heavily on off-policy evaluation or offline model estimation, and can be indirect and inefficient. We argue that a good solution should be able to explicitly parameterize a policy (i.e. respecting action conditionals), implicitly learn from rollout dynamics (i.e. leveraging state marginals), and---crucially---operate in an entirely offline fashion. To address this challenge, we propose a novel technique by energy-based distribution matching (EDM): By identifying parameterizations of the (discriminative) model of a policy with the (generative) energy function for state distributions, EDM yields a simple but effective solution that equivalently minimizes a divergence between the occupancy measure for the demonstrator and a model thereof for the imitator. Through experiments with application to control and healthcare settings, we illustrate consistent performance gains over existing algorithms for strictly batch imitation learning.
**********

On the Ergodicity, Bias and Asymptotic Normality of Randomized Midpoint Sampling Method
Ye He, Krishnakumar Balasubramanian, Murat A. Erdogdu
The randomized midpoint method, proposed by (Shen and Lee, 2019), has emerged as an optimal discretization procedure for simulating the continuous time underdamped Langevin diffusion. In this paper, we analyze several probabilistic properties of the randomized midpoint discretization method, considering both overdamped and underdamped Langevin dynamics. We first characterize the stationary distribution of the discrete chain obtained with constant step-size discretization and show that it is biased away from the target distribution. Notably, the step-size needs to go to zero to obtain asymptotic unbiasedness. Next, we establish the asymptotic normality of numerical integration using the randomized midpoint method and highlight the relative advantages and disadvantages over other discretizations. Our results collectively provide several insights into the behavior of the randomized midpoint discretization method, including obtaining confidence intervals for numerical integrations.
**********
A Single-Loop Smoothed Gradient Descent-Ascent Algorithm for Nonconvex-Concave Min-Max Problems
Jiawei Zhang, Peijun Xiao, Ruoyu Sun, Zhiquan 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.
**********
Generating Correct Answers for Progressive Matrices Intelligence Tests
Niv Pekar, Yaniv Benny, Lior Wolf
Raven’s Progressive Matrices are multiple-choice intelligence tests, where one tries to complete the missing location in a 3x3 grid of abstract images. Previous attempts to address this test have focused solely on selecting the right answer out of the multiple choices. In this work, we focus, instead, on generating a correct answer given the grid, which is a harder task, by definition. The proposed neural model combines multiple advances in generative models, including employing multiple pathways through the same network, using the reparameterization trick along two pathways to make their encoding compatible, a selective application of variational losses, and a complex perceptual loss that is coupled with a selective backpropagation procedure. Our algorithm is able not only to generate a set of plausible answers but also to be competitive to the state of the art methods in multiple-choice tests.
**********
HyNet: Learning Local Descriptor with Hybrid Similarity Measure and Triplet Loss
Yurun Tian, Axel Barroso Laguna, Tony Ng, Vassileios Balntas, Krystian Mikolajczyk
In this paper, we investigate how L2 normalisation affects the back-propagated descriptor gradients during training. Based on our observations, we propose HyNet, a new local descriptor that leads to state-of-the-art results in matching. HyNet introduces a hybrid similarity measure for triplet margin loss, a regularisation term constraining the descriptor norm, and a new network architecture that performs L2 normalisation of all intermediate feature maps and the output descriptors. HyNet surpasses previous methods by a significant margin on standard benchmarks that include patch matching, verification, and retrieval, as well as outperforming full end-to-end methods on 3D reconstruction tasks.
**********
Preference learning along multiple criteria: A game-theoretic perspective
Kush Bhatia, Ashwin Pananjady, Peter Bartlett, Anca Dragan, Martin J. Wainwright
The literature on ranking from ordinal data is vast, and there are several ways to aggregate overall preferences from pairwise comparisons between objects. In particular, it is well-known that any Nash equilibrium of the zero-sum game induced by the preference matrix defines a natural solution concept (winning distribution over objects) known as a von Neumann winner. Many real-world problems, howe

ver, are inevitably multi-criteria, with different pairwise preferences governin g the different criteria. In this work, we generalize the notion of a von Neuman n winner to the multi-criteria setting by taking inspiration from Blackwell’s ap proachability. Our framework allows for non-linear aggregation of preferences ac ross criteria, and generalizes the linearization-based approach from multi-objec tive optimization.
**********

Multi-Plane Program Induction with 3D Box Priors
Yikai Li, Jiayuan Mao, Xiuming Zhang, Bill Freeman, Josh Tenenbaum, Noah Snavely , Jiajun Wu
We consider two important aspects in understanding and editing images: modeling regular, program-like texture or patterns in 2D planes, and 3D posing of these p lanes in the scene. Unlike prior work on image-based program synthesis, which as sumes the image contains a single visible 2D plane, we present Box Program Induc tion (BPI), which infers a program-like scene representation that simultaneously models repeated structure on multiple 2D planes, the 3D position and orientatio n of the planes, and camera parameters, all from a single image. Our model assum es a box prior, i.e., that the image captures either an inner view or an outer v iew of a box in 3D. It uses neural networks to infer visual cues such as vanishi ng points, wireframe lines to guide a search-based algorithm to find the program that best explains the image. Such a holistic, structured scene representation enables 3D-aware interactive image editing operations such as inpainting missing pixels, changing camera parameters, and extrapolate the image contents.
**********

Online Neural Connectivity Estimation with Noisy Group Testing
Anne Draelos, John Pearson
One of the primary goals of systems neuroscience is to relate the structure of n eural circuits to their function, yet patterns of connectivity are difficult to establish when recording from large populations in behaving organisms. Many prev ious approaches have attempted to estimate functional connectivity between neuro ns using statistical modeling of observational data, but these approaches rely h eavily on parametric assumptions and are purely correlational. Recently, however , holographic photostimulation techniques have made it possible to precisely tar get selected ensembles of neurons, offering the possibility of establishing dire ct causal links. A naive method for inferring functional connections is to stimu late each individual neuron multiple times and observe the responses of cells in the local network, but this approach scales poorly with the number of neurons. Here, we propose a method based on noisy group testing that drastically increase s the efficiency of this process in sparse networks. By stimulating small ensemb les of neurons, we show that it is possible to recover binarized network connect ivity with a number of tests that grows only logarithmically with population siz e under minimal statistical assumptions. Moreover, we prove that our approach, w hich reduces to an efficiently solvable convex optimization problem, can be rela ted to Variational Bayesian inference on the binary connection weights, and we d erive rigorous bounds on the posterior marginals. This allows us to extend our m ethod to the streaming setting, where continuously updated posteriors allow for optional stopping, and we demonstrate the feasibility of inferring connectivity for networks of up to tens of thousands of neurons online.
**********

Once-for-All Adversarial Training: In-Situ Tradeoff between Robustness and Accur acy for Free
Haotao Wang, Tianlong Chen, Shupeng Gui, TingKuei Hu, Ji Liu, Zhangyang Wang
Adversarial training and its many variants substantially improve deep network ro bustness, yet at the cost of compromising standard accuracy. Moreover, the train ing process is heavy and hence it becomes impractical to thoroughly explore the trade-off between accuracy and robustness. This paper asks this new question: ho w to quickly calibrate a trained model in-situ, to examine the achievable trade- offs between its standard and robust accuracies, without (re-)training it many t imes? Our proposed framework, Once-for-all Adversarial Training (OAT), is built on an innovative model-conditional training framework, with a controlling hyper-

parameter as the input. The trained model could be adjusted among different stan
dard and robust accuracies “for free” at testing time. As an important knob, we
exploit dual batch normalization to separate standard and adversarial feature st
atistics, so that they can be learned in one model without degrading performance
. We further extend OAT to a Once-for-all Adversarial Training and Slimming (OAT
S) framework, that allows for the joint trade-off among accuracy, robustness and
 runtime efficiency. Experiments show that, without any re-training nor ensembli
ng, OAT/OATS achieve similar or even superior performance compared to dedicatedl
y trained models at various configurations. Our codes and pretrained models are
available at: https://github.com/VITA-Group/Once-for-All-Adversarial-Training.
**********

Implicit Neural Representations with Periodic Activation Functions
Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, Gordon Wetzst
ein
Implicitly defined, continuous, differentiable signal representations parameteri
zed by neural networks have emerged as a powerful paradigm, offering many possib
le benefits over conventional representations. However, current network architec
tures for such implicit neural representations are incapable of modeling signals
 with fine detail, and fail to represent a signal's spatial and temporal derivat
ives, despite the fact that these are essential to many physical signals defined
 implicitly as the solution to partial differential equations. We propose to lev
erage periodic activation functions for implicit neural representations and demo
nstrate that these networks, dubbed sinusoidal representation networks or SIRENs
, are ideally suited for representing complex natural signals and their derivati
ves. We analyze SIREN activation statistics to propose a principled initializati
on scheme and demonstrate the representation of images, wavefields, video, sound
, and their derivatives. Further, we show how SIRENs can be leveraged to solve c
hallenging boundary value problems, such as particular Eikonal equations (yieldi
ng signed distance functions), the Poisson equation, and the Helmholtz and wave
equations. Lastly, we combine SIRENs with hypernetworks to learn priors over the
 space of SIREN functions.
**********

Rotated Binary Neural Network
Mingbao Lin, Rongrong Ji, Zihan Xu, Baochang Zhang, Yan Wang, Yongjian Wu, Feiyu
e Huang, Chia-Wen Lin
Binary Neural Network (BNN) shows its predominance in reducing the complexity of
 deep neural networks. However, it suffers severe performance degradation. One o
f the major impediments is the large quantization error between the full-precisi
on weight vector and its binary vector. Previous works focus on compensating for
 the norm gap while leaving the angular bias hardly touched. In this paper, for
the first time, we explore the influence of angular bias on the quantization err
or and then introduce a Rotated Binary Neural Network (RBNN), which considers th
e angle alignment between the full-precision weight vector and its binarized ver
sion. At the beginning of each training epoch, we propose to rotate the full-pre
cision weight vector to its binary vector to reduce the angular bias. To avoid t
he high complexity of learning a large rotation matrix, we further introduce a b
i-rotation formulation that learns two smaller rotation matrices. In the trainin
g stage, we devise an adjustable rotated weight vector for binarization to escap
e the potential local optimum. Our rotation leads to around 50% weight flips whi
ch maximize the information gain. Finally, we propose a training-aware approxima
tion of the sign function for the gradient backward. Experiments on CIFAR-10 and
 ImageNet demonstrate the superiorities of RBNN over many state-of-the-arts. Our
 source code, experimental settings, training logs and binary models are availab
le at https://github.com/lmbxmu/RBNN.
**********

Community detection in sparse time-evolving graphs with a dynamical Bethe-Hessia
n
Lorenzo Dall'Amico, Romain Couillet, Nicolas Tremblay
This article considers the problem of community detection in sparse dynamical gr
aphs in which the community structure evolves over time. A fast spectral algorit

hm based on an extension of the Bethe-Hessian matrix is proposed, which benefits from the positive correlation in the class labels and in their temporal evolution and is designed to be applicable to any dynamical graph with a community structure. Under the dynamical degree-corrected stochastic block model, in the case of two classes of equal size, we demonstrate and support with extensive simulations that our proposed algorithm is capable of making non-trivial community reconstruction as soon as theoretically possible, thereby reaching the optimal detectability threshold and provably outperforming competing spectral methods.
**********

Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness
Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, Balaji Lakshminarayanan
Bayesian neural networks (BNN) and deep ensembles are principled approaches to estimate the predictive uncertainty of a deep learning model. However their practicality in real-time, industrial-scale applications are limited due to their heavy memory and inference cost. This motivates us to study principled approaches to high-quality uncertainty estimation that require only a single deep neural network (DNN). By formalizing the uncertainty quantification as a minimax learning problem, we first identify input distance awareness, i.e., the model’s ability to quantify the distance of a testing example from the training data in the input space, as a necessary condition for a DNN to achieve high-quality (i.e., minimax optimal) uncertainty estimation. We then propose Spectral-normalized Neural Gaussian Process (SNGP), a simple method that improves the distance-awareness ability of modern DNNs, by adding a weight normalization step during training and replacing the output layer. On a suite of vision and language understanding tasks and on modern architectures (Wide-ResNet and BERT), SNGP is competitive with deep ensembles in prediction, calibration and out-of-domain detection, and outperforms the other single-model approaches.
**********

Adaptive Learning of Rank-One Models for Efficient Pairwise Sequence Alignment
Govinda Kamath, Tavor Baharav, Ilan Shomorony
Pairwise alignment of DNA sequencing data is a ubiquitous task in bioinformatics and typically represents a heavy computational burden. State-of-the-art approaches to speed up this task use hashing to identify short segments (k-mers) that are shared by pairs of reads, which can then be used to estimate alignment scores. However, when the number of reads is large, accurately estimating alignment scores for all pairs is still very costly. Moreover, in practice, one is only interested in identifying pairs of reads with large alignment scores. In this work, we propose a new approach to pairwise alignment estimation based on two key new ingredients. The first ingredient is to cast the problem of pairwise alignment estimation under a general framework of rank-one crowdsourcing models, where the workers' responses correspond to k-mer hash collisions. These models can be accurately solved via a spectral decomposition of the response matrix. The second ingredient is to utilise a multi-armed bandit algorithm to adaptively refine this spectral estimator only for read pairs that are likely to have large alignments. The resulting algorithm iteratively performs a spectral decomposition of the response matrix for adaptively chosen subsets of the read pairs.
**********

Hierarchical nucleation in deep neural networks
Diego Doimo, Aldo Glielmo, Alessio Ansuini, Alessandro Laio
Deep convolutional networks (DCNs) learn meaningful representations where data that share the same abstract characteristics are positioned closer and closer. Understanding these representations and how they are generated is of unquestioned practical and theoretical interest. In this work we study the evolution of the probability density of the ImageNet dataset across the hidden layers in some state-of-the-art DCNs.
We find that the initial layers generate a unimodal probability density getting rid of any structure irrelevant for classification. In subsequent layers density peaks arise in a hierarchical fashion that mirrors the semantic hierarchy of th

e concepts. Density peaks corresponding to single categories appear only close t
o the output and via a very sharp transition which resembles the nucleation proc
ess of a heterogeneous liquid. This process leaves a  footprint in the probabili
ty density of the output layer where the topography of the  peaks allows reconst
ructing the semantic relationships of the categories.
**********

Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional
Domains
Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin R
aghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, Ren Ng
We show that passing input points through a simple Fourier feature mapping enabl
es a multilayer perceptron (MLP) to learn high-frequency functions in low-dimens
ional problem domains. These results shed light on recent advances in computer v
ision and graphics that achieve state-of-the-art results by using MLPs to repres
ent complex 3D objects and scenes. Using tools from the neural tangent kernel (N
TK) literature, we show that a standard MLP has impractically slow convergence t
o high frequency signal components. To overcome this spectral bias, we use a Fou
rier feature mapping to transform the effective NTK into a stationary kernel wit
h a tunable bandwidth. We suggest an approach for selecting problem-specific Fou
rier features that greatly improves the performance of MLPs for low-dimensional
regression tasks relevant to the computer vision and graphics communities.
**********

Graph Geometry Interaction Learning
Shichao Zhu, Shirui Pan, Chuan Zhou, Jia Wu, Yanan Cao, Bin Wang
While numerous approaches have been developed to embed graphs into either Euclid
ean or hyperbolic spaces, they do not fully utilize the information available in
 graphs, or lack the flexibility to model intrinsic complex graph geometry. To u
tilize the strength of both Euclidean and hyperbolic geometries, we develop a no
vel Geometry Interaction Learning (GIL) method for graphs, a well-suited and eff
icient alternative for learning abundant geometric properties in graph. GIL capt
ures a more informative internal structural features with low dimensions while m
aintaining conformal invariance of each space. Furthermore, our method endows ea
ch node the freedom to determine the importance of each geometry space via a fle
xible dual feature interaction learning and probability assembling mechanism. Pr
omising experimental results are presented for five benchmark datasets on node c
lassification and link prediction tasks.
**********

Differentiable Augmentation for Data-Efficient GAN Training
Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, Song Han
The performance of generative adversarial networks (GANs) heavily deteriorates g
iven a limited amount of training data. This is mainly because the discriminator
sis memorizing the exact training set. To combat it, we propose Differentiable A
ugmentation (DiffAugment), a simple method that improves the data efficiency of
GANs by imposing various types of differentiable augmentations on both real and
fake samples. Previous attempts to directly augment the training data manipulate
 the distribution of real images, yielding little benefit; DiffAugment enables u
s to adopt the differentiable augmentation for the generated samples, effectivel
y stabilizes training, and leads to better convergence. Experiments demonstrate
consistent gains of our method over a variety of GAN architectures and loss func
tions for both unconditional and class-conditional generation. With DiffAugment,
 we achieve astate-of-the-art FID of 6.80 with an IS of 100.8 on ImageNet 128×12
8 and 2-4× reductions of FID given 1,000 images on FFHQ and LSUN. Furthermore, w
ith only 20% training data, we can match the top performance on CIFAR-10 and CIF
AR-100. Finally, our method can generate high-fidelity images using only 100 ima
ges without pre-training, while being on par with existing transfer learning alg
orithms. Code is available at https://github.com/mit-han-lab/data-efficient-gans
.
**********

Heuristic Domain Adaptation
Shuhao Cui, Xuan Jin, Shuhui Wang, Yuan He, Qingming Huang

In visual domain adaptation (DA), separating the domain-specific characteristics from the domain-invariant representations is an ill-posed problem. Existing methods apply different kinds of priors or directly minimize the domain discrepancy to address this problem, which lack flexibility in handling real-world situations. Another research pipeline expresses the domain-specific information as a gradual transferring process, which tends to be suboptimal in accurately removing the domain-specific properties. In this paper, we address the modeling of domain-invariant and domain-specific information from the heuristic search perspective. We identify the characteristics in the existing representations that lead to larger domain discrepancy as the heuristic representations. With the guidance of heuristic representations, we formulate a principled framework of Heuristic Domain Adaptation (HDA) with well-founded theoretical guarantees. To perform HDA, the cosine similarity scores and independence measurements between domain-invariant and domain-specific representations are cast into the constraints at the initial and final states during the learning procedure. Similar to the final condition of heuristic search, we further derive a constraint enforcing the final range of heuristic network output to be small. Accordingly, we propose Heuristic Domain Adaptation Network (HDAN), which explicitly learns the domain-invariant and domain-specific representations with the above mentioned constraints. Extensive experiments show that HDAN has exceeded state-of-the-art on unsupervised DA, multi-source DA and semi-supervised DA. The code is available at https://github.com/cuishuhao/HDA.
**********
Learning Certified Individually Fair Representations
Anian Ruoss, Mislav Balunovic, Marc Fischer, Martin Vechev
Fair representation learning provides an effective way of enforcing fairness constraints without compromising utility for downstream users. A desirable family of such fairness constraints, each requiring similar treatment for similar individuals, is known as individual fairness. In this work, we introduce the first method that enables data consumers to obtain certificates of individual fairness for existing and new data points. The key idea is to map similar individuals to close latent representations and leverage this latent proximity to certify individual fairness. That is, our method enables the data producer to learn and certify a representation where for a data point all similar individuals are at l-infinity distance at most epsilon, thus allowing data consumers to certify individual fairness by proving epsilon-robustness of their classifier. Our experimental evaluation on five real-world datasets and several fairness constraints demonstrates the expressivity and scalability of our approach.
**********
Part-dependent Label Noise: Towards Instance-dependent Label Noise
Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, Masashi Sugiyama
Learning with the \textit{instance-dependent} label noise is challenging, because it is hard to model such real-world noise. Note that there are psychological and physiological evidences showing that we humans perceive instances by decomposing them into parts. Annotators are therefore more likely to annotate instances based on the parts rather than the whole instances, where a wrong mapping from parts to classes may cause the instance-dependent label noise. Motivated by this human cognition, in this paper, we approximate the instance-dependent label noise by exploiting \textit{part-dependent} label noise. Specifically, since instances can be approximately reconstructed by a combination of parts, we approximate the instance-dependent \textit{transition matrix} for an instance by a combination of the transition matrices for the parts of the instance. The transition matrices for parts can be learned by exploiting anchor points (i.e., data points that belong to a specific class almost surely). Empirical evaluations on synthetic and real-world datasets demonstrate our method is superior to the state-of-the-art approaches for learning from the instance-dependent label noise.
**********
Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, H. Vincent Poor
In federated learning, heterogeneity in the clients' local datasets and computation speeds results in large variations in the number of local updates performed by each client in each communication round. Naive weighted aggregation of such models causes objective inconsistency, that is, the global model converges to a stationary point of a mismatched objective function which can be arbitrarily different from the true objective. This paper provides a general framework to analyze the convergence of federated heterogeneous optimization algorithms. It subsumes previously proposed methods such as FedAvg and FedProx and provides the first principled understanding of the solution bias and the convergence slowdown due to objective inconsistency. Using insights from this analysis, we propose FedNova, a normalized averaging method that eliminates objective inconsistency while preserving fast error convergence.
**********

An Improved Analysis of (Variance-Reduced) Policy Gradient and Natural Policy Gradient Methods
Yanli Liu, Kaiqing Zhang, Tamer Basar, Wotao Yin
In this paper, we revisit and improve the convergence of policy gradient (PG), natural PG (NPG) methods, and their variance-reduced variants, under general smooth policy parametrizations. More specifically, with the Fisher information matrix of the policy being positive definite: i) we show that a state-of-the-art variance-reduced PG method, which has only been shown to converge to stationary points, converges to the globally optimal value up to some inherent function approximation error due to policy parametrization; ii) we show that NPG enjoys a lower sample complexity; iii) we propose SRVR-NPG, which incorporates variance-reduction into the NPG update. Our improvements follow from an observation that the convergence of (variance-reduced) PG and NPG methods can improve each other: the stationary convergence analysis of PG can be applied on NPG as well, and the global convergence analysis of NPG can help to establish the global convergence of (variance-reduced) PG methods. Our analysis carefully integrates the advantages of these two lines of works. Thanks to this improvement, we have also made variance-reduction for NPG possible for the first time, with both global convergence and an efficient finite-sample complexity.
**********

Geometric Exploration for Online Control
Orestis Plevrakis, Elad 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.
**********

Automatic Curriculum Learning through Value Disagreement
Yunzhi Zhang, Pieter Abbeel, Lerrel Pinto
Continually solving new, unsolved tasks is the key to learning diverse behaviors. Through reinforcement learning (RL), we have made massive strides towards solving tasks that have a single goal. However, in the multi-task domain, where an agent needs to reach multiple goals, the choice of training goals can largely affect sample efficiency. When biological agents learn, there is often an organized and meaningful order to which learning happens. Inspired by this, we propose setting up an automatic curriculum for goals that the agent needs to solve. Our key insight is that if we can sample goals at the frontier of the set of goals that an agent is able to reach, it will provide a significantly stronger learning signal compared to randomly sampled goals. To operationalize this idea, we introduce a goal proposal module that prioritizes goals that maximize the epistemic uncertainty of the Q-function of the policy. This simple technique samples goals that are neither too hard nor too easy for the agent to solve, hence enabling continual improvement. We evaluate our method across 13 multi-goal robotic tasks and 5 navigation tasks, and demonstrate performance gains over current state-of-the-art methods.
**********

MRI Banding Removal via Adversarial Training
Aaron Defazio, Tullie Murrell, Michael Recht
MR images reconstructed from sub-sampled Cartesian data using deep learning techniques show a characteristic banding (sometimes described as streaking), which is particularly strong in low signal-to-noise regions of the reconstructed image. In this work, we propose the use of an adversarial loss that penalizes banding structures without requiring any human annotation. Our technique greatly reduces the appearance of banding, without requiring any additional computation or post-processing at reconstruction time. We report the results of a blind comparison against a strong baseline by a group of expert evaluators (board-certified radiologists), where our approach is ranked superior at banding removal with no statistically significant loss of detail. A reference implementation of our method is available in the supplementary material.
**********

The NetHack Learning Environment
Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, Tim Rocktäschel
Progress in Reinforcement Learning (RL) algorithms goes hand-in-hand with the development of challenging environments that test the limits of current methods. While existing RL environments are either sufficiently complex or based on fast simulation, they are rarely both. Here, we present the NetHack Learning Environment (NLE), a scalable, procedurally generated, stochastic, rich, and challenging environment for RL research based on the popular single-player terminal-based roguelike game, NetHack. We argue that NetHack is sufficiently complex to drive long-term research on problems such as exploration, planning, skill acquisition, and language-conditioned RL, while dramatically reducing the computational resources required to gather a large amount of experience. We compare NLE and its task suite to existing alternatives, and discuss why it is an ideal medium for testing the robustness and systematic generalization of RL agents. We demonstrate empirical success for early stages of the game using a distributed Deep RL baseline and Random Network Distillation exploration, alongside qualitative analysis of various agents trained in the environment. NLE is open source and available at https://github.com/facebookresearch/nle.
**********

Language and Visual Entity Relationship Graph for Agent Navigation
Yicong Hong, Cristian Rodriguez, Yuankai Qi, Qi Wu, Stephen Gould
Vision-and-Language Navigation (VLN) requires an agent to navigate in a real-world environment following natural language instructions. From both the textual and visual perspectives, we find that the relationships among the scene, its objects, and directional cues are essential for the agent to interpret complex instructions and correctly perceive the environment. To capture and utilize the relationships, we propose a novel Language and Visual Entity Relationship Graph for modelling the inter-modal relationships between text and vision, and the intra-modal relationships among visual entities. We propose a message passing algorithm for propagating information between language elements and visual entities in the graph, which we then combine to determine the next action to take. Experiments show that by taking advantage of the relationships we are able to improve over state-of-the-art. On the Room-to-Room (R2R) benchmark, our method achieves the new best performance on the test unseen split with success rate weighted by path length of 52%. On the Room-for-Room (R4R) dataset, our method significantly improves the previous best from 13% to 34% on the success weighted by normalized dynamic time warping.
**********

ICAM: Interpretable Classification via Disentangled Representations and Feature Attribution Mapping
Cher Bass, Mariana da Silva, Carole Sudre, Petru-Daniel Tudosiu, Stephen Smith, Emma Robinson
Feature attribution (FA), or the assignment of class-relevance to different locations in an image, is important for many classification problems but is particularly crucial within the neuroscience domain, where accurate mechanistic models o

f behaviours, or disease, require knowledge of all features discriminative of a trait. At the same time, predicting class relevance from brain images is challenging as phenotypes are typically heterogeneous, and changes occur against a background of significant natural variation.  Here, we present a novel framework for creating class specific FA maps through image-to-image translation. We propose the use of a VAE-GAN to explicitly disentangle class relevance from background features for improved interpretability properties, which results in meaningful FA maps. We validate our method on 2D and 3D brain image datasets of dementia (ADNI dataset), ageing (UK Biobank), and (simulated) lesion detection. We show that FA maps generated by our method outperform baseline FA methods when validated against ground truth. More significantly, our approach is the first to use latent space sampling to support exploration of phenotype variation.
**********

Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks

Zhou Fan, Zhichao Wang

We study the eigenvalue distributions of the Conjugate Kernel and Neural Tangent Kernel associated to multi-layer feedforward neural networks. In an asymptotic regime where network width is increasing linearly in sample size, under random initialization of the weights, and for input samples satisfying a notion of approximate pairwise orthogonality, we show that the eigenvalue distributions of the CK and NTK converge to deterministic limits. The limit for the CK is described by iterating the Marcenko-Pastur map across the hidden layers. The limit for the NTK is equivalent to that of a linear combination of the CK matrices across layers, and may be described by recursive fixed-point equations that extend this Marcenko-Pastur map. We demonstrate the agreement of these asymptotic predictions with the observed spectra for both synthetic and CIFAR-10 training data, and we perform a small simulation to investigate the evolutions of these spectra over training.
**********

No-Regret Learning Dynamics for Extensive-Form Correlated Equilibrium

Andrea Celli, Alberto Marchesi, Gabriele Farina, Nicola Gatti

The existence of simple, uncoupled no-regret dynamics that converge to correlated equilibria in normal-form games is a celebrated result in the theory of multi-agent systems. Specifically, it has been known for more than 20 years that when all players seek to minimize their internal regret in a repeated normal-form game, the empirical frequency of play converges to a normal-form correlated equilibrium. Extensive-form (that is, tree-form) games generalize normal-form games by modeling both sequential and simultaneous moves, as well as private information. Because of the sequential nature and presence of partial information in the game, extensive-form correlation has significantly different properties than the normal-form counterpart, many of which are still open research directions. Extensive-form correlated equilibrium (EFCE) has been proposed as the natural extensive-form counterpart to normal-form correlated equilibrium. However, it was currently unknown whether EFCE emerges as the result of uncoupled agent dynamics. In this paper, we give the first uncoupled no-regret dynamics that converge to the set of EFCEs in n-player general-sum extensive-form games with perfect recall. First, we introduce a notion of trigger regret in extensive-form games, which extends that of internal regret in normal-form games. When each player has low trigger regret, the empirical frequency of play is a close to an EFCE. Then, we give an efficient no-trigger-regret algorithm. Our algorithm decomposes trigger regret into local subproblems at each decision point for the player, and constructs a global strategy of the player from the local solutions at each decision point.
**********

Estimating weighted areas under the ROC curve

Andreas Maurer, Massimiliano Pontil

Exponential bounds on the estimation error are given for the plug-in estimator of weighted areas under the ROC curve. The bounds hold for single score functions and uniformly over classes of functions, whose complexity can be controlled by Gaussian or Rademacher averages. The results justify learning algorithms which s

elect score functions to maximize the empirical partial area under the curve (pAUC). They also illustrate the use of some recent advances in the theory of nonlinear empirical processes.
**********

Can Implicit Bias Explain Generalization? Stochastic Convex Optimization as a Case Study
Assaf Dauber, Meir Feder, Tomer Koren, Roi Livni
The notion of implicit bias, or implicit regularization, has been suggested as a means to explain the surprising generalization ability of modern-days overparameterized learning algorithms. This notion refers to the tendency of the optimization algorithm towards a certain structured solution that often generalizes well. Recently, several papers have studied implicit regularization and were able to identify this phenomenon in various scenarios.
**********

Generalized Hindsight for Reinforcement Learning
Alexander Li, Lerrel Pinto, Pieter Abbeel
One of the key reasons for the high sample complexity in reinforcement learning (RL) is the inability to transfer knowledge from one task to another. In standard multi-task RL settings, low-reward data collected while trying to solve one task provides little to no signal for solving that particular task and is hence effectively wasted. However, we argue that this data, which is uninformative for one task, is likely a rich source of information for other tasks. To leverage this insight and efficiently reuse data, we present Generalized Hindsight: an approximate inverse reinforcement learning technique for relabeling behaviors with the right tasks. Intuitively, given a behavior generated under one task, Generalized Hindsight returns a different task that the behavior is better suited for. Then, the behavior is relabeled with this new task before being used by an off-policy RL optimizer. Compared to standard relabeling techniques, Generalized Hindsight provides a substantially more efficient re-use of samples, which we empirically demonstrate on a suite of multi-task navigation and manipulation tasks.
**********

Critic Regularized Regression
Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S. Merel, Jost Tobias Springenberg, Scott E. Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, Nando de Freitas
Offline reinforcement learning (RL), also known as batch RL, offers the prospect of policy optimization from large pre-recorded datasets without online environment interaction. It addresses challenges with regard to the cost of data collection and safety, both of which are particularly pertinent to real-world applications of RL. Unfortunately, most off-policy algorithms perform poorly when learning from a fixed dataset. In this paper, we propose a novel offline RL algorithm to learn policies from data using a form of critic-regularized regression (CRR). We find that CRR performs surprisingly well and scales to tasks with high-dimensional state and action spaces -- outperforming several state-of-the-art offline RL algorithms by a significant margin on a wide range of benchmark tasks.
**********

Boosting Adversarial Training with Hypersphere Embedding
Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, Hang Su
Adversarial training (AT) is one of the most effective defenses against adversarial attacks for deep learning models. In this work, we advocate incorporating the hypersphere embedding (HE) mechanism into the AT procedure by regularizing the features onto compact manifolds, which constitutes a lightweight yet effective module to blend in the strength of representation learning. Our extensive analyses reveal that AT and HE are well coupled to benefit the robustness of the adversarially trained models from several aspects. We validate the effectiveness and adaptability of HE by embedding it into the popular AT frameworks including PGD-AT, ALP, and TRADES, as well as the FreeAT and FastAT strategies. In the experiments, we evaluate our methods under a wide range of adversarial attacks on the CIFAR-10 and ImageNet datasets, which verifies that integrating HE can consistently enhance the model robustness for each AT framework with little extra computat

ion.
**********
Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Des
igns
Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, Danai Koutra
We investigate the representation power of graph neural networks in the semi-sup
ervised node classification task under heterophily or low homophily, i.e., in ne
tworks where connected nodes may have different class labels and dissimilar feat
ures. Many popular GNNs fail to generalize to this setting, and are even outperf
ormed by models that ignore the graph structure (e.g., multilayer perceptrons).
Motivated by this limitation, we identify a set of key designs—ego- and neighbor
-embedding separation, higher-order neighborhoods, and combination of intermedia
te representations—that boost learning from the graph structure under heterophil
y. We combine them into a graph neural network, H2GCN, which we use as the base
method to empirically evaluate the effectiveness of the identified designs. Goin
g beyond the traditional benchmarks with strong homophily, our empirical analysi
s shows that the identified designs increase the accuracy of GNNs by up to 40% a
nd 27% over models without them on synthetic and real networks with heterophily,
 respectively, and yield competitive performance under homophily.
**********
Modeling Continuous Stochastic Processes with Dynamic Normalizing Flows
Ruizhi Deng, Bo Chang, Marcus A. Brubaker, Greg Mori, Andreas Lehrmann
Normalizing flows transform a simple base distribution into a complex target dis
tribution and have proved to be powerful models for data generation and density
estimation. In this work, we propose a novel type of normalizing flow driven by
a differential deformation of the continuous-time Wiener process. As a result, w
e obtain a rich time series model whose observable process inherits many of the
appealing properties of its base process, such as efficient computation of likel
ihoods and marginals. Furthermore, our continuous treatment provides a natural f
ramework for irregular time series with an independent arrival process, includin
g straightforward interpolation. We illustrate the desirable properties of the p
roposed model on popular stochastic processes and demonstrate its superior flexi
bility to variational RNN and latent ODE baselines in a series of experiments on
 synthetic and real-world data.
**********
Efficient Online Learning of Optimal Rankings: Dimensionality Reduction via Grad
ient Descent
Dimitris Fotakis, Thanasis Lianeas, Georgios Piliouras, Stratis Skoulakis
We consider a natural model of online preference aggregation, where sets of pref
erred items R1, R2, ..., Rt, ..., along with a demand for kt items in each Rt, a
ppear online. Without prior knowledge of (Rt, kt), the learner maintains a ranki
ng \pit aiming that at least kt items from Rt appear high in \pi_t. This is a fu
ndamental problem in preference aggregation with applications to e.g., ordering
product or news items in web pages based on user scrolling and click patterns.
**********
Training Normalizing Flows with the Information Bottleneck for Competitive Gener
ative Classification
Lynton Ardizzone, Radek Mackowiak, Carsten Rother, Ullrich Köthe
The Information Bottleneck (IB) objective uses information theory to formulate a
 task-performance versus robustness trade-off. It has been successfully applied
in the standard discriminative classification setting. We pose the question whet
her the IB can also be used to train generative likelihood models such as normal
izing flows. Since normalizing flows use invertible network architectures (INNs)
, they are information-preserving by construction. This seems contradictory to t
he idea of a bottleneck. In this work, firstly, we develop the theory and method
ology of IB-INNs, a class of conditional normalizing flows where INNs are traine
d using the IB objective: Introducing a small amount of controlled information l
oss allows for an asymptotically exact formulation of the IB, while keeping the
INN's generative capabilities intact. Secondly, we investigate the properties of
 these models experimentally, specifically used as generative classifiers. This

model class offers advantages such as improved uncertainty quantification and ou
t-of-distribution detection, but traditional generative classifier solutions suf
fer considerably in classification accuracy. We find the trade-off parameter in
the IB controls a mix of generative capabilities and accuracy close to standard
classifiers. Empirically, our uncertainty estimates in this mixed regime compare
 favourably to conventional generative and discriminative classifiers. Code is p
rovided in the supplement.
**********
Detecting Hands and Recognizing Physical Contact in the Wild
Supreeth Narasimhaswamy, Trung Nguyen, Minh Hoai Nguyen
We investigate a new problem of detecting hands and recognizing their physical c
ontact state in unconstrained conditions. This is a challenging inference task g
iven the need to reason beyond the local appearance of hands. The lack of traini
ng annotations indicating which object or parts of an object the hand is in cont
act with further complicates the task. We propose a novel convolutional network
based on Mask-RCNN that can jointly learn to localize hands and predict their ph
ysical contact to address this problem.  The network uses outputs from another o
bject detector to obtain locations of objects present in the scene. It uses thes
e outputs and hand locations to recognize the hand's contact state using two att
ention mechanisms. The first attention mechanism is based on the hand and a regi
on's affinity, enclosing the hand and the object, and densely pools features fro
m this region to the hand region. The second attention module adaptively selects
 salient features from this plausible region of contact. To develop and evaluate
 our method's performance, we introduce a large-scale dataset called ContactHand
s, containing unconstrained images annotated with hand locations and contact sta
tes. The proposed network, including the parameters of attention modules, is end
-to-end trainable. This network achieves approximately 7% relative improvement o
ver a baseline network that was built on the vanilla Mask-RCNN architecture and
trained for recognizing hand contact states.
**********
On the Theory of Transfer Learning: The Importance of Task Diversity
Nilesh Tripuraneni, Michael Jordan, Chi Jin
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Finite-Time Analysis of Round-Robin Kullback-Leibler Upper Confidence Bounds for
 Optimal Adaptive Allocation with Multiple Plays and Markovian Rewards
Vrettos Moulos
We study an extension of the classic stochastic multi-armed bandit problem which
 involves multiple plays and Markovian rewards in the rested bandits setting.
In order to tackle this problem we consider an adaptive allocation rule which at
 each stage combines the information from the sample means of all the arms,
with the Kullback-Leibler upper confidence bound of a single arm which is select
ed in round-robin way.
For rewards generated from a one-parameter exponential family of Markov chains,
we provide a finite-time upper bound for the regret incurred from this adaptive
allocation rule, which reveals the logarithmic dependence of the regret on the t
ime horizon, and which is asymptotically optimal.
For our analysis we devise several concentration results for Markov chains,
including a maximal inequality for Markov chains,
that may be of interest in their own right.
As a byproduct of our analysis we also establish asymptotically optimal, finite-
time guarantees for the case of multiple plays, and i.i.d. rewards drawn from a
one-parameter exponential family of probability densities.
Additionally, we provide simulation results that illustrate that calculating Kul
lback-Leibler upper confidence bounds in a round-robin way, is significantly mor
e efficient than calculating them for every arm at each round, and that the expe
cted regrets of those two approaches behave similarly.

**********
Neural Star Domain as Primitive Representation
Yuki Kawana, Yusuke Mukuta, Tatsuya Harada
Reconstructing 3D objects from 2D images is a fundamental task in computer visio n. Acurate structured reconstruction by parsimonious and semantic primitive repr esentation further broadens its application. When reconstructing a target shape with multiple primitives, it is preferable that one can instantly access the uni on of basic properties of the shape such as collective volume and surface, treat ing the primitives as if they are one single shape. This becomes possible by pri mitive representation with unified implicit and explicit representations. Howeve r, primitive representations in current approaches do not satisfy all of the abo ve requirements at the same time. To solve this problem, we propose a novel prim itive representation named neural star domain (NSD) that learns primitive shapes in the star domain. We show that NSD is a universal approximator of the star do main and is not only parsimonious and semantic but also an implicit and explicit shape representation. We demonstrate that our approach outperforms existing met hods in image reconstruction tasks, semantic capabilities, and speed and quality of sampling high-resolution meshes.
**********
Off-Policy Interval Estimation with Lipschitz Value Iteration
Ziyang Tang, Yihao Feng, Na Zhang, Jian Peng, Qiang Liu
Off-policy evaluation provides an essential tool for evaluating the effects of d ifferent policies or treatments using only observed data. When applied to high-s takes scenarios such as medical diagnosis or financial decision-making, it is es sential to provide provably correct upper and lower bounds of the expected rewar d, not just a classical single point estimate, to the end-users, as executing a poor policy can be very costly. In this work, we propose a provably correct meth od for obtaining interval bounds for off-policy evaluation in a general continuo us setting. The idea is to search for the maximum and minimum values of the expe cted reward among all the Lipschitz Q-functions that are consistent with the obs ervations, which amounts to solving a constrained optimization problem on a Lips chitz function space. We go on to introduce a Lipschitz value iteration method t o monotonically tighten the interval, which is simple yet efficient and provably convergent. We demonstrate the practical efficiency of our method on a range of benchmarks.
**********
Inverse Rational Control with Partially Observable Continuous Nonlinear Dynamics
Minhae Kwon, Saurabh Daptardar, Paul R. Schrater, Xaq Pitkow
A fundamental question in neuroscience is how the brain creates an internal mode l of the world to guide actions using sequences of ambiguous sensory information . This is naturally formulated as a reinforcement learning problem under partial observations, where an agent must estimate relevant latent variables in the wor ld from its evidence, anticipate possible future states, and choose actions that optimize total expected reward. This problem can be solved by control theory, w hich allows us to find the optimal actions for a given system dynamics and objec tive function. However, animals often appear to behave suboptimally. Why? We hyp othesize that animals have their own flawed internal model of the world, and cho ose actions with the highest expected subjective reward according to that flawed model. We describe this behavior as {\it rational} but not optimal. The problem of Inverse Rational Control (IRC) aims to identify which internal model would b est explain an agent's actions. Our contribution here generalizes past work on I nverse Rational Control which solved this problem for discrete control in partia lly observable Markov decision processes. Here we accommodate continuous nonline ar dynamics and continuous actions, and impute sensory observations corrupted by unknown noise that is private to the animal. We first build an optimal Bayesian agent that learns an optimal policy generalized over the entire model space of dynamics and subjective rewards using deep reinforcement learning. Crucially, t his allows us to compute a likelihood over models for experimentally observable action trajectories acquired from a suboptimal agent. We then find the model par ameters that maximize the likelihood using gradient ascent. Our method successfu

lly recovers the true model of rational agents. This approach provides a foundation for interpreting the behavioral and neural dynamics of animal brains during complex tasks.

**********

Deep Statistical Solvers

Balthazar Donon, Zhengying Liu, Wenzhuo LIU, Isabelle Guyon, Antoine Marot, Marc Schoenauer

This paper introduces Deep Statistical Solvers (DSS), a new class of trainable solvers for optimization problems, arising e.g., from system simulations. The key idea is to learn a solver that generalizes to a given distribution of problem instances. This is achieved by directly using as loss the objective function of the problem, as opposed to most previous Machine Learning based approaches, which mimic the solutions attained by an existing solver. Though both types of approaches outperform classical solvers with respect to speed for a given accuracy, a distinctive advantage of DSS is that they can be trained without a training set of sample solutions. Focusing on use cases of systems of interacting and interchangeable entities (e.g. molecular dynamics, power systems, discretized PDEs), the proposed approach is instantiated within a class of Graph Neural Networks. Under sufficient conditions, we prove that the corresponding set of functions contains approximations to any arbitrary precision of the actual solution of the optimization problem. The proposed approach is experimentally validated on large linear problems, demonstrating super-generalisation properties; And on AC power grid simulations, on which the predictions of the trained model have a correlation higher than 99.99% with the outputs of the classical Newton-Raphson method (known for its accuracy), while being 2 to 3 orders of magnitude faster.

**********

Distributionally Robust Parametric Maximum Likelihood Estimation

Viet Anh Nguyen, Xuhui Zhang, Jose Blanchet, Angelos Georghiou

We consider the parameter estimation problem of a probabilistic generative model prescribed using a natural exponential family of distributions. For this problem, the typical maximum likelihood estimator usually overfits under limited training sample size, is sensitive to noise and may perform poorly on downstream predictive tasks. To mitigate these issues, we propose a distributionally robust maximum likelihood estimator that minimizes the worst-case expected log-loss uniformly over a parametric Kullback-Leibler ball around a parametric nominal distribution. Leveraging the analytical expression of the Kullback-Leibler divergence between two distributions in the same natural exponential family, we show that the min-max estimation problem is tractable in a broad setting, including the robust training of generalized linear models. Our novel robust estimator also enjoys statistical consistency and delivers promising empirical results in both regression and classification tasks.

**********

Secretary and Online Matching Problems with Machine Learned Advice

Antonios Antoniadis, Themis Gouleakis, Pieter Kleer, Pavel Kolev

The classical analysis of online algorithms, due to its worst-case nature, can be quite pessimistic when the input instance at hand is far from worst-case. Often this is not an issue with machine learning approaches, which shine in exploiting patterns in past inputs in order to predict the future. However, such predictions, although usually accurate, can be arbitrarily poor. Inspired by a recent line of work, we augment three well-known online settings with machine learned predictions about the future, and develop algorithms that take them into account. In particular, we study the following online selection problems: (i) the classical secretary problem, (ii) online bipartite matching and (iii) the graphic matroid secretary problem. Our algorithms still come with a worst-case performance guarantee in the case that predictions are subpar while obtaining an improved competitive ratio (over the best-known classical online algorithm for each problem) when the predictions are sufficiently accurate. For each algorithm, we establish a trade-off between the competitive ratios obtained in the two respective cases.

**********

Deep Transformation-Invariant Clustering
Tom Monnier, Thibault Groueix, Mathieu Aubry
Recent advances in image clustering typically focus on learning better deep repr
esentations. In contrast, we present an orthogonal approach that does not rely o
n abstract features but instead learns to predict transformations and performs c
lustering directly in image space. This learning process naturally fits in the g
radient-based training of K-means and Gaussian mixture model, without requiring
any additional loss or hyper-parameters. It leads us to two new deep transformat
ion-invariant clustering frameworks, which jointly learn prototypes and transfor
mations. More specifically, we use deep learning modules that enable us to resol
ve invariance to spatial, color and morphological transformations. Our approach
is conceptually simple and comes with several advantages, including the possibil
ity to easily adapt the desired invariance to the task and a strong interpretabi
lity of both cluster centers and assignments to clusters.  We demonstrate that o
ur novel approach yields competitive and highly promising results on standard im
age clustering benchmarks. Finally, we showcase its robustness and the advantage
s of its improved interpretability by visualizing clustering results over real p
hotograph collections.
**********
Overfitting Can Be Harmless for Basis Pursuit, But Only to a Degree
Peizhong Ju, Xiaojun Lin, Jia Liu
Recently, there have been significant interests in studying the so-called "doubl
e-descent" of the generalization error of linear regression models under the ove
rparameterized and overfitting regime, with the hope that such analysis may prov
ide the first step towards understanding why overparameterized deep neural netwo
rks (DNN) still generalize well. However, to date most of these studies focused
on the min L2-norm solution that overfits the data. In contrast, in this paper w
e study the overfitting solution that minimizes the L1-norm, which is known as B
asis Pursuit (BP) in the compressed sensing literature. Under a sparse true line
ar regression model with p i.i.d. Gaussian features, we show that for a large ra
nge of p up to a limit that grows exponentially with the number of samples n, wi
th high probability the model error of BP is upper bounded by a value that decre
ases with p. To the best of our knowledge, this is the first analytical result i
n the literature establishing the double-descent of overfitting BP for finite n
and p. Further, our results reveal significant differences between the double-de
scent of BP and min L2-norm solutions. Specifically, the double-descent upper-bo
und of BP is independent of the signal strength, and for high SNR and sparse mod
els the descent-floor of BP can be much lower and wider than that of min L2-norm
 solutions.
**********
Improving Generalization in Reinforcement Learning with Mixture Regularization
KAIXIN WANG, Bingyi Kang, Jie Shao, Jiashi Feng
Deep reinforcement learning (RL) agents trained in a limited set of environments
 tend to suffer overfitting and fail to generalize to unseen testing environment
s. To improve their generalizability, data augmentation approaches (e.g. cutout
and random convolution) are previously explored to increase the data diversity.
However, we find these approaches only locally perturb the observations regardle
ss of the training environments, showing limited effectiveness on enhancing the
data diversity and the generalization performance. In this work, we introduce a
simple approach, named mixreg, which trains agents on a mixture of observations
from different training environments and imposes linearity constraints on the ob
servation interpolations and the supervision (e.g. associated reward) interpolat
ions. Mixreg increases the data diversity more effectively and helps learn smoot
her policies. We verify its effectiveness on improving generalization by conduct
ing extensive experiments on the large-scale Procgen benchmark. Results show mix
reg outperforms the well-established baselines on unseen testing environments by
 a large margin. Mixreg is simple, effective and general. It can be applied to b
oth policy-based and value-based RL algorithms. Code is available at https://git
hub.com/kaixin96/mixreg.
**********

Pontryagin Differentiable Programming: An End-to-End Learning and Control Framework
Wanxin Jin, Zhaoran Wang, Zhuoran Yang, Shaoshuai Mou
This paper develops a Pontryagin differentiable programming (PDP) methodology, which establishes a unified framework to solve a broad class of learning and control tasks. The PDP distinguishes from existing methods by two novel techniques: first, we differentiate through Pontryagin's Maximum Principle, and this allows to obtain the analytical derivative of a trajectory with respect to tunable parameters within an optimal control system, enabling end-to-end learning of dynamics, policies, or/and control objective functions; and second, we propose an auxiliary control system in the backward pass of the PDP framework, and the output of this auxiliary control system is the analytical derivative of the original system's trajectory with respect to the parameters, which can be iteratively solved using standard control tools. We investigate three learning modes of the PDP: inverse reinforcement learning, system identification, and control/planning. We demonstrate the capability of the PDP in each learning mode on different high-dimensional systems, including multilink robot arm, 6-DoF maneuvering UAV, and 6-DoF rocket powered landing.
**********
Learning from Aggregate Observations
Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, Masashi Sugiyama
We study the problem of learning from aggregate observations where supervision signals are given to sets of instances instead of individual instances, while the goal is still to predict labels of unseen individuals. A well-known example is multiple instance learning (MIL). In this paper, we extend MIL beyond binary classification to other problems such as multiclass classification and regression. We present a general probabilistic framework that accommodates a variety of aggregate observations, e.g., pairwise similarity/triplet comparison for classification and mean/difference/rank observation for regression. Simple maximum likelihood solutions can be applied to various differentiable models such as deep neural networks and gradient boosting machines. Moreover, we develop the concept of consistency up to an equivalence relation to characterize our estimator and show that it has nice convergence properties under mild assumptions. Experiments on three problem settings --- classification via triplet comparison and regression via mean/rank observation indicate the effectiveness of the proposed method.
**********
The Devil is in the Detail: A Framework for Macroscopic Prediction via Microscopic Models
Yingxiang Yang, Negar Kiyavash, Le Song, Niao He
Macroscopic data aggregated from microscopic events are pervasive in machine learning, such as country-level COVID-19 infection statistics based on city-level data. Yet, many existing approaches for predicting macroscopic behavior only use aggregated data, leaving a large amount of fine-grained microscopic information unused. In this paper, we propose a principled optimization framework for macroscopic prediction by fitting microscopic models based on conditional stochastic optimization. The framework leverages both macroscopic and microscopic information, and adapts to individual microscopic models involved in the aggregation. In addition, we propose efficient learning algorithms with convergence guarantees. In our experiments, we show that the proposed learning framework clearly outperforms other plug-in supervised learning approaches in real-world applications, including the prediction of daily infections of COVID-19 and medicare claims.
**********
Subgraph Neural Networks
Emily Alsentzer, Samuel Finlayson, Michelle Li, Marinka Zitnik
Deep learning methods for graphs achieve remarkable performance on many node-level and graph-level prediction tasks. However, despite the proliferation of the methods and their success, prevailing Graph Neural Networks (GNNs) neglect subgraphs, rendering subgraph prediction tasks challenging to tackle in many impactful applications. Further, subgraph prediction tasks present several unique challenges: subgraphs can have non-trivial internal topology, but also carry a notion o

f position and external connectivity information relative to the underlying grap h in which they exist. Here, we introduce SubGNN, a subgraph neural network to l earn disentangled subgraph representations. We propose a novel subgraph routing mechanism that propagates neural messages between the subgraph’s components and randomly sampled anchor patches from the underlying graph, yielding highly accur ate subgraph representations. SubGNN specifies three channels, each designed to capture a distinct aspect of subgraph topology, and we provide empirical evidenc e that the channels encode their intended properties. We design a series of new synthetic and real-world subgraph datasets. Empirical results for subgraph class ification on eight datasets show that SubGNN achieves considerable performance g ains, outperforming strong baseline methods, including node-level and graph-leve l GNNs, by 19.8% over the strongest baseline. SubGNN performs exceptionally well on challenging biomedical datasets, where subgraphs have complex topology and e ven comprise multiple disconnected components.
**********

Demystifying Orthogonal Monte Carlo and Beyond
Han Lin, Haoxian Chen, Krzysztof M. Choromanski, Tianyi Zhang, Clement Laroche
Orthogonal Monte Carlo (OMC) is a very effective sampling algorithm imposing str uctural geometric conditions (orthogonality) on samples for variance reduction. Due to its simplicity and superior performance as compared to its Quasi Monte Ca rlo counterparts, OMC is used in a wide spectrum of challenging machine learning applications ranging from scalable kernel methods to predictive recurrent neura l networks, generative models and reinforcement learning.
However theoretical understanding of the method remains very limited. In this pa per we shed new light on the theoretical principles behind OMC, applying theory of negatively dependent random variables to obtain several new concentration res ults.
As a corollary, we manage to obtain first uniform convergence results for OMCs a nd consequently, substantially strengthen best known downstream guarantees for k ernel ridge regression via OMCs. We also propose novel extensions of the method leveraging theory of algebraic varieties over finite fields and particle algorit hms, called Near-Orthogonal Monte Carlo (NOMC). We show that NOMC is the first a lgorithm consistently outperforming OMC in applications ranging from kernel meth ods to approximating distances in probabilistic metric spaces.
**********

Optimal Robustness-Consistency Trade-offs for Learning-Augmented Online Algorith ms
Alexander Wei, Fred Zhang
We study the problem of improving the performance of online algorithms by incorp orating machine-learned predictions. The goal is to design algorithms that are b oth consistent and robust, meaning that the algorithm performs well when predict ions are accurate and maintains worst-case guarantees. Such algorithms have been studied in a recent line of works due to Lykouris and Vassilvitskii (ICML '18) and Purohit et al (NeurIPS '18). They provide robustness-consistency trade-offs for a variety of online problems. However, they leave open the question of whet her these trade-offs are tight, i.e., to what extent to such trade-offs are nece ssary. In this paper, we provide the first set of non-trivial lower bounds for competitive analysis using machine-learned predictions. We focus on the classic problems of ski-rental and non-clairvoyant scheduling and provide optimal trade-offs in various settings.
**********

A Scalable Approach for Privacy-Preserving Collaborative Machine Learning
Jinhyun So, Basak Guler, Salman Avestimehr
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search
Jaehyeon Kim, Sungwon Kim, Jungil Kong, Sungroh Yoon

Recently, text-to-speech (TTS) models such as FastSpeech and ParaNet have been p
roposed to generate mel-spectrograms from text in parallel. Despite the advantag
e, the parallel TTS models cannot be trained without guidance from autoregressiv
e TTS models as their external aligners. In this work, we propose Glow-TTS, a fl
ow-based generative model for parallel TTS that does not require any external al
igner. By combining the properties of flows and dynamic programming, the propose
d model searches for the most probable monotonic alignment between text and the
latent representation of speech on its own. We demonstrate that enforcing hard m
onotonic alignments enables robust TTS, which generalizes to long utterances, an
d employing generative flows enables fast, diverse, and controllable speech synt
hesis. Glow-TTS obtains an order-of-magnitude speed-up over the autoregressive m
odel, Tacotron 2, at synthesis with comparable speech quality. We further show t
hat our model can be easily extended to a multi-speaker setting.
**********

Towards Learning Convolutions from Scratch
Behnam Neyshabur
Convolution is one of the most essential components of modern architectures used
 in computer vision. As machine learning moves towards reducing the expert bias
and learning it from data, a natural next step seems to be learning convolution-
like structures from scratch.  This, however, has proven elusive.  For example,
current state-of-the-art architecture search algorithms use convolution as one o
f the existing modules rather than learning it from data. In an attempt to under
stand the inductive bias that gives rise to convolutions, we investigate minimum
 description length as a guiding principle and show that in some settings, it ca
n indeed be indicative of the performance of architectures. To find architecture
s with small description length, we propose beta-LASSO, a simple variant of LASS
O algorithm that, when applied on fully-connected networks for image classificat
ion tasks, learns architectures with local connections and achieves state-of-the
-art accuracies for training fully-connected networks on CIFAR-10 (84.50%), CIFA
R-100 (57.76%) and SVHN (93.84%) bridging the gap between fully-connected and co
nvolutional networks.
**********

Cycle-Contrast for Self-Supervised Video Representation Learning
Quan Kong, Wenpeng Wei, Ziwei Deng, Tomoaki Yoshinaga, Tomokazu Murakami
We present Cycle-Contrastive Learning (CCL), a novel self-supervised method for
learning video representation. Following a nature that there is a belong and inc
lusion relation of video and its frames, CCL is designed to find correspondences
 across frames and videos considering the contrastive representation in their do
mains respectively. It is different from recent approaches that merely learn cor
respondences across frames or clips. In our method, the frame and video represen
tations are learned from a single network based on an R3D network, with a shared
 non-linear transformation for embedding both frame and video features before th
e cycle-contrastive loss. We demonstrate that the video representation learned b
y CCL can be transferred well to downstream tasks of video understanding, outper
forming previous methods in nearest neighbour retrieval and action recognition t
asks on UCF101, HMDB51 and MMAct.
**********

Posterior Re-calibration for Imbalanced Datasets
Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, Zsolt Kira
Neural Networks can perform poorly when the training label distribution is heavi
ly imbalanced, as well as when the testing data differs from the training distri
bution. In order to deal with shift in the testing label distribution, which imb
alance causes, we motivate the problem from the perspective of an optimal Bayes
classifier and derive a prior rebalancing technique that can be solved through a
 KL-divergence based optimization. This method allows a flexible post-training h
yper-parameter to be efficiently tuned on a validation set and effectively modif
y the classifier margin to deal with this imbalance. We further combine this met
hod with existing likelihood shift methods, re-interpreting them from the same B
ayesian perspective, and demonstrating that our method can deal with both proble
ms in a unified way. The resulting algorithm can be conveniently used on probabi

listic classification problems agnostic to underlying architectures. Our results on six different datasets and five different architectures show state of art accuracy, including on large-scale imbalanced datasets such as iNaturalist for classification and Synthia for semantic segmentation. Please see https://github.com/GT-RIPL/UNO-IC.git for implementation.
**********

Novelty Search in Representational Space for Sample Efficient Exploration
Ruo Yu Tao, Vincent Francois-Lavet, Joelle Pineau
We present a new approach for efficient exploration which leverages a low-dimensional encoding of the environment learned with a combination of model-based and model-free objectives.
Our approach uses intrinsic rewards that are based on the distance of nearest neighbors in the low dimensional representational space to gauge novelty.
We then leverage these intrinsic rewards for sample-efficient exploration with planning routines in representational space for hard exploration tasks with sparse rewards.
One key element of our approach is the use of information theoretic principles to shape our representations in a way so that our novelty reward goes beyond pixel similarity.
We test our approach on a number of maze tasks, as well as a control problem and show that our exploration approach is more sample-efficient compared to strong baselines.
**********

Robust Reinforcement Learning via Adversarial training with Langevin Dynamics
Parameswaran Kamalaruban, Yu-Ting Huang, Ya-Ping Hsieh, Paul Rolland, Cheng Shi, Volkan Cevher
We introduce a \emph{sampling} perspective to tackle the challenging task of training robust Reinforcement Learning (RL) agents. Leveraging the powerful Stochastic Gradient Langevin Dynamics, we present a novel, scalable two-player RL algorithm, which is a sampling variant of the two-player policy gradient method. Our algorithm consistently outperforms existing baselines, in terms of generalization across different training and testing conditions, on several MuJoCo environments. Our experiments also show that, even for objective functions that entirely ignore potential environmental shifts, our sampling approach remains highly robust in comparison to standard RL algorithms.
**********

Adversarial Blocking Bandits
Nicholas Bishop, Hau Chan, Debmalya Mandal, Long Tran-Thanh
We consider a general adversarial multi-armed blocking bandit setting where each played arm can be blocked (unavailable) for some time periods and the reward per arm is given at each time period adversarially without obeying any distribution. The setting models scenarios of allocating scarce limited supplies (e.g., arms) where the supplies replenish and can be reused only after certain time periods. We first show that, in the optimization setting, when the blocking durations and rewards are known in advance, finding an optimal policy (e.g., determining which arm per round) that maximises the cumulative reward is strongly NP-hard, eliminating the possibility of a fully polynomial-time approximation scheme (FPTAS) for the problem unless P = NP.  To complement our result, we show that a greedy algorithm that plays the best available arm at each round provides an approximation guarantee that depends on the blocking durations and the path variance of the rewards. In the bandit setting, when the blocking durations and rewards are not known, we design two algorithms, RGA and RGA-META, for the case of bounded duration an path variation. In particular, when the variation budget BT is known in advance, RGA can achieve O(\sqrt{T(2\tilde{D}+K)B{T}}) dynamic approximate regret. On the other hand, when B_T is not known, we show that the dynamic approximate regret of RGA-META is at most O((K+\tilde{D})^{1/4}\tilde{B}^{1/2}T^{3/4}) where \tilde{B} is the maximal path variation budget within each batch of RGA-META (which is provably in order of o(\sqrt{T}). We also prove that if either the variation budget or the maximal blocking duration is unbounded, the approximate regret will be at least Theta(T). We also show that the regret upper bound of  R

GA is tight if the blocking durations are bounded above by an order of O(1).
**********
Online Algorithms for Multi-shop Ski Rental with Machine Learned Advice
Shufan Wang, Jian Li, Shiqiang Wang
We study the problem of augmenting online algorithms with machine learned (ML) advice. In particular, we consider the \emph{multi-shop ski rental} (MSSR) problem, which is a generalization of the classical ski rental problem. In MSSR, each shop has different prices for buying and renting a pair of skis, and a skier has to make decisions on when and where to buy. We obtain both deterministic and randomized online algorithms with provably improved performance when either a single or multiple ML predictions are used to make decisions. These online algorithms have no knowledge about the quality or the prediction error type of the ML prediction. The performance of these online algorithms are robust to the poor performance of the predictors, but improve with better predictions. Extensive experiments using both synthetic and real world data traces verify our theoretical observations and show better performance against algorithms that purely rely on online decision making.
**********
Multi-label Contrastive Predictive Coding
Jiaming Song, Stefano Ermon
Variational mutual information (MI) estimators are widely used in unsupervised representation learning methods such as contrastive predictive coding (CPC). A lower bound on MI can be obtained from a multi-class classification problem, where a critic attempts to distinguish a positive sample drawn from the underlying joint distribution from (m-1) negative samples drawn from a suitable proposal distribution. Using this approach, MI estimates are bounded above by \log m, and could thus severely underestimate unless m is very large. To overcome this limitation, we introduce a novel estimator based on a multi-label classification problem, where the critic needs to jointly identify \emph{multiple} positive samples at the same time. We show that using the same amount of negative samples, multi-label CPC is able to exceed the \log m bound, while still being a valid lower bound of mutual information. We demonstrate that the proposed approach is able to lead to better mutual information estimation, gain empirical improvements in unsupervised representation learning, and beat the current state-of-the-art in knowledge distillation over 10 out of 13 tasks.
**********
Rotation-Invariant Local-to-Global Representation Learning for 3D Point Cloud
SEOHYUN KIM, JaeYoo Park, Bohyung Han
We propose a local-to-global representation learning algorithm for 3D point cloud data, which is appropriate to handle various geometric transformations, especially rotation, without explicit data augmentation with respect to the transformations.
Our model takes advantage of multi-level abstraction based on graph convolutional neural networks, which constructs a descriptor hierarchy to encode rotation-invariant shape information of an input object in a bottom-up manner.
The descriptors in each level are obtained from a neural network based on a graph via stochastic sampling of 3D points, which is effective in making the learned representations robust to the variations of input data.
The proposed algorithm presents the state-of-the-art performance on the rotation-augmented 3D object recognition and segmentation benchmarks, and we further analyze its characteristics through comprehensive ablative experiments.
**********
Learning Invariants through Soft Unification
Nuri Cingillioglu, Alessandra Russo
Human reasoning involves recognising common underlying principles across many examples. The by-products of such reasoning are invariants that capture patterns such as "if someone went somewhere then they are there", expressed using variables "someone" and "somewhere" instead of mentioning specific people or places. Humans learn what variables are and how to use them at a young age. This paper explores whether machines can also learn and use variables solely from examples with

out requiring human pre-engineering. We propose Unification Networks, an end-to-end differentiable neural network approach capable of lifting examples into invariants and using those invariants to solve a given task. The core characteristic of our architecture is soft unification between examples that enables the network to generalise parts of the input into variables, thereby learning invariants. We evaluate our approach on five datasets to demonstrate that learning invariants captures patterns in the data and can improve performance over baselines.
**********

One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL

Saurabh Kumar, Aviral Kumar, Sergey Levine, Chelsea Finn

While reinforcement learning algorithms can learn effective policies for complex tasks, these policies are often brittle to even minor task variations, especially when variations are not explicitly provided during training. One natural approach to this problem is to train agents with manually specified variation in the training task or environment. However, this may be infeasible in practical situations, either because making perturbations is not possible, or because it is unclear how to choose suitable perturbation strategies without sacrificing performance. The key insight of this work is that learning diverse behaviors for accomplishing a task can directly lead to behavior that generalizes to varying environments, without needing to perform explicit perturbations during training. By identifying multiple solutions for the task in a single environment during training, our approach can generalize to new situations by abandoning solutions that are no longer effective and adopting those that are. We theoretically characterize a robustness set of environments that arises from our algorithm and empirically find that our diversity-driven approach can extrapolate to various changes in the environment and task.
**********

Variational Bayesian Monte Carlo with Noisy Likelihoods

Luigi Acerbi

Variational Bayesian Monte Carlo (VBMC) is a recently introduced framework that uses Gaussian process surrogates to perform approximate Bayesian inference in models with black-box, non-cheap likelihoods. In this work, we extend VBMC to deal with noisy log-likelihood evaluations, such as those arising from simulation-based models. We introduce new global' acquisition functions, such as expected information gain (EIG) and variational interquantile range (VIQR), which are robust to noise and can be efficiently evaluated within the VBMC setting. In a novel, challenging, noisy-inference benchmark comprising of a variety of models with real datasets from computational and cognitive neuroscience, VBMC+VIQR achieves state-of-the-art performance in recovering the ground-truth posteriors and model evidence.

In particular, our method vastly outperformslocal' acquisition functions and other surrogate-based inference methods while keeping a small algorithmic cost. Our benchmark corroborates VBMC as a general-purpose technique for sample-efficient black-box Bayesian inference also with noisy models.
**********

Finite-Sample Analysis of Contractive Stochastic Approximation Using Smooth Convex Envelopes

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, Karthikeyan Shanmugam

Stochastic Approximation (SA) is a popular approach for solving fixed-point equations where the information is corrupted by noise. In this paper, we consider an SA involving a contraction mapping with respect to an arbitrary norm, and show its finite-sample error bounds while using different stepsizes. The idea is to construct a smooth Lyapunov function using the generalized Moreau envelope, and show that the iterates of SA have negative drift with respect to that Lyapunov function.  Our result is applicable in Reinforcement Learning (RL). In particular, we use it to establish the first-known convergence rate of the V-trace algorithm for off-policy TD-learning [18]. Importantly, our construction results in only a logarithmic dependence of the convergence bound on the size of the state-space.

**********
Self-Supervised Generative Adversarial Compression
Chong Yu, Jeff Pool
Deep learning’s success has led to larger and larger models to handle more and m ore complex tasks; trained models often contain millions of parameters. These la rge models are compute- and memory-intensive, which makes it a challenge to depl oy them with latency, throughput, and storage constraints. Some model compressio n methods have been successfully applied to image classification and detection o r language models, but there has been very little work compressing generative ad versarial networks (GANs) performing complex tasks. In this paper, we show that a standard model compression technique, weight pruning and knowledge distillatio n, cannot be applied to GANs using existing methods. We then develop a self-supe rvised compression technique which uses the trained discriminator to supervise t he training of a compressed generator. We show that this framework has compellin g performance to high degrees of sparsity, can be easily applied to new tasks an d models, and enables meaningful comparisons between different compression granu larities.
**********
An efficient nonconvex reformulation of stagewise convex optimization problems
Rudy R. Bunel, Oliver Hinder, Srinadh Bhojanapalli, Krishnamurthy Dvijotham
Convex optimization problems with staged structure appear in several contexts, i ncluding optimal control, verification of deep neural networks, and isotonic reg ression. Off-the-shelf solvers can solve these problems but may scale poorly. We develop a nonconvex reformulation designed to exploit this staged structure. Ou r reformulation has only simple bound constraints, enabling solution via project ed gradient methods and their accelerated variants. The method automatically gen erates a sequence of primal and dual feasible solutions to the original convex p roblem, making optimality certification easy. We establish theoretical propertie s of the nonconvex formulation, showing that it is (almost) free of spurious loc al minima and has the same global optimum as the convex problem. We modify proje cted gradient descent to avoid spurious local minimizers so it always converges to the global minimizer. For neural network verification, our approach obtains s mall duality gaps in only a few gradient steps. Consequently, it can provide ti ght duality gaps for many large-scale verification problems where both off-the-s helf and specialized solvers struggle.
**********
From Finite to Countable-Armed Bandits
Anand Kalvit, Assaf Zeevi
We consider a stochastic bandit problem with countably many arms that belong to a finite set of types, each characterized by a unique mean reward. In addition, there is a fixed distribution over types which sets the proportion of each type in the population of arms. The decision maker is oblivious to the type of any ar m and to the aforementioned distribution over types, but perfectly knows the tot al number of types occurring in the population of arms. We propose a fully adapt ive online learning algorithm that achieves O(log n) distribution-dependent expe cted cumulative regret after any number of plays n, and show that this order of regret is best possible. The analysis of our algorithm relies on newly discovere d concentration and convergence properties of optimism-based policies like UCB i n finite-armed bandit problems with zero gap, which may be of independent intere st.
**********
Adversarial Distributional Training for Robust Deep Learning
Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, Hang Su
Adversarial training (AT) is among the most effective techniques to improve mode l robustness by augmenting training data with adversarial examples. However, mos t existing AT methods adopt a specific attack to craft adversarial examples, lea ding to the unreliable robustness against other unseen attacks. Besides, a singl e attack algorithm could be insufficient to explore the space of perturbations. In this paper, we introduce adversarial distributional training (ADT), a novel f ramework for learning robust models. ADT is formulated as a minimax optimization

problem, where the inner maximization aims to learn an adversarial distribution to characterize the potential adversarial examples around a natural one under an entropic regularizer, and the outer minimization aims to train robust models by minimizing the expected loss over the worst-case adversarial distributions. Through a theoretical analysis, we develop a general algorithm for solving ADT, and present three approaches for parameterizing the adversarial distributions, ranging from the typical Gaussian distributions to the flexible implicit ones. Empirical results on several benchmarks validate the effectiveness of ADT compared with the state-of-the-art AT methods.
**********

Meta-Learning Stationary Stochastic Process Prediction with Convolutional Neural Processes
Andrew Foong, Wessel Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, Richard Turner
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Theory-Inspired Path-Regularized Differential Network Architecture Search
Pan Zhou, Caiming Xiong, Richard Socher, Steven Chu Hong Hoi
Despite its high search efficiency, differential architecture search (DARTS) often selects network architectures with dominated skip connections which lead to performance degradation. However, theoretical understandings on this issue remain absent, hindering the development of more advanced methods in a principled way. In this work, we solve this problem by theoretically analyzing the effects of various types of operations, e.g. convolution, skip connection and zero operation, to the network optimization. We prove that the architectures with more skip connections can converge faster than the other candidates, and thus are selected by DARTS. This result, for the first time, theoretically and explicitly reveals the impact of skip connections to fast network optimization and its competitive advantage over other types of operations in DARTS. Then we propose a theory-inspired path-regularized DARTS that consists of two key modules: (i) a differential group-structured sparse binary gate introduced for each operation to avoid unfair competition among operations, and (ii) a path-depth-wise regularization used to incite search exploration for deep architectures that often converge slower than shallow ones as shown in our theory and are not well explored during search. Experimental results on image classification tasks validate its advantages. Codes and models will be released.
**********

Conic Descent and its Application to Memory-efficient Optimization over Positive Semidefinite Matrices
John C. Duchi, Oliver Hinder, Andrew Naber, Yinyu Ye
We present an extension of the conditional gradient method to problems whose feasible sets are convex cones. We provide a convergence analysis for the method and for variants with nonconvex objectives, and we extend the analysis to practical cases with effective line search strategies. For the specific case of the positive semidefinite cone, we present a memory-efficient version based on randomized matrix sketches and advocate a heuristic greedy step that greatly improves its practical performance. Numerical results on phase retrieval and matrix completion problems indicate that our method can offer substantial advantages over traditional conditional gradient and Burer-Monteiro approaches.
**********

Learning the Geometry of Wave-Based Imaging
Konik Kothari, Maarten de Hoop, Ivan Dokmani■
We propose a general physics-based deep learning architecture for wave-based imaging problems. A key difficulty in imaging problems with a varying background wave speed is that the medium ``bends'' the waves differently depending on their position and direction. This space-bending geometry makes the equivariance to translations of convolutional networks an undesired inductive bias. We build an int

erpretable neural architecture inspired by Fourier integral operators (FIOs) whi
ch approximate the wave physics. FIOs model a wide range of imaging modalities,
from seismology and radar to Doppler and ultrasound. We focus on learning the ge
ometry of wave propagation captured by FIOs, which is implicit in the data, via
a loss based on optimal transport. The proposed FIONet performs significantly be
tter than the usual baselines on a number of imaging inverse problems, especiall
y in out-of-distribution tests.
**********

Greedy inference with structure-exploiting lazy maps
Michael Brennan, Daniele Bigoni, Olivier Zahm, Alessio Spantini, Youssef Marzouk
We propose a framework for solving high-dimensional Bayesian inference problems
using \emph{structure-exploiting} low-dimensional transport maps or flows. These
 maps are confined to a low-dimensional subspace (hence, lazy), and the subspace
 is identified by minimizing an upper bound on the Kullback--Leibler divergence
(hence, structured). Our framework provides a principled way of identifying and
exploiting low-dimensional structure in an inference problem. It focuses the exp
ressiveness of a transport map along the directions of most significant discrepa
ncy from the posterior, and can be used to build deep compositions of lazy maps,
 where low-dimensional projections of the parameters are iteratively transformed
 to match the posterior. We prove weak convergence of the generated sequence of
distributions to the posterior, and we demonstrate the benefits of the framework
 on challenging inference problems in machine learning and differential equation
s, using inverse autoregressive flows and polynomial maps as examples of the und
erlying density estimators.
**********

Nimble: Lightweight and Parallel GPU Task Scheduling for Deep Learning
Woosuk Kwon, Gyeong-In Yu, Eunji Jeong, Byung-Gon Chun
Deep learning (DL) frameworks take advantage of GPUs to improve the speed of DL
inference and training. Ideally, DL frameworks should be able to fully utilize t
he computation power of GPUs such that the running time depends on the amount of
 computation assigned to GPUs. Yet, we observe that in scheduling GPU tasks, exi
sting DL frameworks suffer from inefficiencies such as large scheduling overhead
 and unnecessary serial execution. To this end, we propose Nimble, a DL executio
n engine that runs GPU tasks in parallel with minimal scheduling overhead. Nimbl
e introduces a novel technique called ahead-of-time (AoT) scheduling. Here, the
scheduling procedure finishes before executing the GPU kernel, thereby removing
most of the scheduling overhead during run time. Furthermore, Nimble automatical
ly parallelizes the execution of GPU tasks by exploiting multiple GPU streams in
 a single GPU. Evaluation on a variety of neural networks shows that compared to
 PyTorch, Nimble speeds up inference and training by up to 22.34× and 3.61×, res
pectively. Moreover, Nimble outperforms state-of-the-art inference systems, Tens
orRT and TVM, by up to 2.81× and 1.70×, respectively.
**********

Finding the Homology of Decision Boundaries with Active Learning
Weizhi Li, Gautam Dasarathy, Karthikeyan Natesan Ramamurthy, Visar Berisha
Accurately and efficiently characterizing the decision boundary of classifiers i
s important for problems related to model selection and meta-learning. Inspired
by topological data analysis, the characterization of decision boundaries using
their homology has recently emerged as a general and powerful tool. In this pape
r, we propose an active learning algorithm to recover the homology of decision b
oundaries. Our algorithm sequentially and adaptively selects which samples it re
quires the labels of.  We theoretically analyze the proposed framework and show
that the query complexity of our active learning algorithm depends naturally on
the intrinsic complexity of the underlying manifold. We demonstrate the effectiv
eness of our framework in selecting best-performing machine learning models for
datasets just using their respective homological summaries. Experiments on sever
al standard datasets show the sample complexity improvement in recovering the ho
mology and demonstrate the practical utility of the framework for model selectio
n.
**********

Reinforced Molecular Optimization with Neighborhood-Controlled Grammars
Chencheng Xu, Qiao Liu, Minlie Huang, Tao Jiang
A major challenge in the pharmaceutical industry is to design novel molecules wi th specific desired properties, especially when the property evaluation is costl y. Here, we propose MNCE-RL, a graph convolutional policy network for molecular optimization with molecular neighborhood-controlled embedding grammars through r einforcement learning. We extend the original neighborhood-controlled embedding grammars to make them applicable to molecular graph generation and design an eff icient algorithm to infer grammatical production rules from given molecules. The use of grammars guarantees the validity of the generated molecular structures. By transforming molecular graphs to parse trees with the inferred grammars, the molecular structure generation task is modeled as a Markov decision process wher e a policy gradient strategy is utilized. In a series of experiments, we demonst rate that our approach achieves state-of-the-art performance in a diverse range of molecular optimization tasks and exhibits significant superiority in optimizi ng molecular properties with a limited number of property evaluations.
**********

Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Proce sses
Dongsheng Ding, Kaiqing Zhang, Tamer Basar, Mihailo Jovanovic
We study sequential decision-making problems in which each agent aims to maximiz e the expected total reward while satisfying a constraint on the expected total utility. We employ the natural policy gradient method to solve the discounted in finite-horizon Constrained Markov Decision Processes (CMDPs) problem. Specifical ly, we propose a new Natural Policy Gradient Primal-Dual (NPG-PD) method for CMD Ps which updates the primal variable via natural policy gradient ascent and the dual variable via projected sub-gradient descent. Even though the underlying max imization involves a nonconcave objective function and a nonconvex constraint se t under the softmax policy parametrization, we prove that our method achieves gl obal convergence with sublinear rates regarding both the optimality gap and the constraint violation. Such a convergence is independent of the size of the state -action space, i.e., it is~dimension-free. Furthermore, for the general smooth p olicy class, we establish sublinear rates of convergence regarding both the opti mality gap and the constraint violation, up to a function approximation error ca used by restricted policy parametrization. Finally, we show that two sample-base d NPG-PD algorithms inherit such non-asymptotic convergence properties and provi de finite-sample complexity guarantees. To the best of our knowledge, our work i s the first to establish non-asymptotic convergence guarantees of policy-based p rimal-dual methods for solving infinite-horizon discounted CMDPs. We also provid e computational results to demonstrate merits of our approach.
**********

Classification Under Misspecification: Halfspaces, Generalized Linear Models, an d Evolvability
Sitan Chen, Frederic Koehler, Ankur Moitra, Morris Yau
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Certified Defense to Image Transformations via Randomized Smoothing
Marc Fischer, Maximilian Baader, Martin Vechev
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
**********

Estimation of Skill Distribution from a Tournament
Ali Jadbabaie, Anuran Makur, Devavrat 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.

**********

Reparameterizing Mirror Descent as Gradient Descent

Ehsan Amid, Manfred K. K. Warmuth

Most of the recent successful applications of neural networks have been based on training with gradient descent updates. However, for some small networks, other mirror descent updates learn provably more efficiently when the target is sparse. We present a general framework for casting a mirror descent update as a gradient descent update on a different set of parameters. In some cases, the mirror descent reparameterization can be described as training a modified network with standard backpropagation. The reparameterization framework is versatile and covers a wide range of mirror descent updates, even cases where the domain is constrained. Our construction for the reparameterization argument is done for the continuous versions of the updates. Finding general criteria for the discrete versions to closely track their continuous counterparts remains an interesting open problem.

**********

General Control Functions for Causal Effect Estimation from IVs

Aahlad Puli, Rajesh Ranganath

Causal effect estimation relies on separating the variation in the outcome into parts due to the treatment and due to the confounders. To achieve this separation, practitioners often use external sources of randomness that only influence the treatment called instrumental variables (IVs). We study variables constructed from treatment and IV that help estimate effects, called control functions. We characterize general control functions for effect estimation in a meta-identification result. Then, we show that structural assumptions on the treatment process allow the construction of general control functions, thereby guaranteeing identification. To construct general control functions and estimate effects, we develop the general control function method (GCFN). GCFN’s first stage called variational decoupling (VDE) constructs general control functions by recovering the residual variation in the treatment given the IV. Using VDE’s control function, GCFN’s second stage estimates effects via regression. Further, we develop semi-supervised GCFN to construct general control functions using subsets of data that have both IV and confounders observed as supervision; this needs no structural treatment process assumptions. We evaluate GCFN on low and high dimensional simulated data and on recovering the causal effect of slave export on modern community trust [30]

**********

Optimal Algorithms for Stochastic Multi-Armed Bandits with Heavy Tailed Rewards

Kyungjae Lee, Hongjun Yang, Sungbin Lim, Songhwai Oh

**********

Certified Robustness of Graph Convolution Networks for Graph Classification under Topological Attacks

Hongwei Jin, Zhan Shi, Venkata Jaya Shankar Ashish Peruri, Xinhua Zhang

Graph convolution networks (GCNs) have become effective models for graph classification. Similar to many deep networks, GCNs are vulnerable to adversarial attacks on graph topology and node attributes. Recently, a number of effective attack and defense algorithms have been designed, but no certificate of robustness has been developed for GCN-based graph classification under topological perturbations with both local and global budgets. In this paper, we propose the first certificate for this problem. Our method is based on Lagrange dualization and convex envelope, which result in tight approximation bounds that are efficiently computable by dynamic programming. When used in conjunction with robust training, it allows an increased number of graphs to be certified as robust.

**********

Zero-Resource Knowledge-Grounded Dialogue Generation

Linxiao Li, Can Xu, Wei Wu, YUFAN ZHAO, Xueliang Zhao, Chongyang Tao

While neural conversation models have shown great potentials towards generating informative and engaging responses via introducing external knowledge, learning such a model often requires knowledge-grounded dialogues that are difficult to obtain. To overcome the data challenge and reduce the cost of building a knowledge-grounded dialogue system, we explore the problem under a zero-resource setting by assuming no context-knowledge-response triples are needed for training. To this end, we propose representing the knowledge that bridges a context and a response and the way that the knowledge is expressed as latent variables, and devise a variational approach that can effectively estimate a generation model from independent dialogue corpora and knowledge corpora. Evaluation results on three benchmarks of knowledge-grounded dialogue generation indicate that our model can achieve comparable performance with state-of-the-art methods that rely on knowledge-grounded dialogues for training, and exhibits a good generalization ability over different datasets.

**********

Targeted Adversarial Perturbations for Monocular Depth Prediction

Alex Wong, Safa Cicek, Stefano Soatto

We study the effect of adversarial perturbations on the task of monocular depth prediction. Specifically, we explore the ability of small, imperceptible additive perturbations to selectively alter the perceived geometry of the scene. We show that such perturbations can not only globally re-scale the predicted distances from the camera, but also alter the prediction to match a different target scene. We also show that, when given semantic or instance information, perturbations can fool the network to alter the depth of specific categories or instances in the scene, and even remove them while preserving the rest of the scene. To understand the effect of targeted perturbations, we conduct experiments on state-of-the-art monocular depth prediction methods. Our experiments reveal vulnerabilities in monocular depth prediction networks, and shed light on the biases and context learned by them.

**********

Beyond the Mean-Field: Structured Deep Gaussian Processes Improve the Predictive Uncertainties

Jakob Lindinger, David Reeb, Christoph Lippert, Barbara Rakitsch

Deep Gaussian Processes learn probabilistic data representations for supervised learning by cascading multiple Gaussian Processes. While this model family promises flexible predictive distributions, exact inference is not tractable. Approximate inference techniques trade off the ability to closely resemble the posterior distribution against speed of convergence and computational efficiency. We propose a novel Gaussian variational family that allows for retaining covariances between latent processes while achieving fast convergence by marginalising out all global latent variables. After providing a proof of how this marginalisation can be done for general covariances, we restrict them to the ones we empirically found to be most important in order to also achieve computational efficiency. We provide an efficient implementation of our new approach and apply it to several benchmark datasets. It yields excellent results and strikes a better balance between accuracy and calibrated uncertainty estimates than its state-of-the-art alternatives.

**********

Offline Imitation Learning with a Misspecified Simulator

Shengyi Jiang, Jingcheng Pang, Yang 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.

**********

Multi-Fidelity Bayesian Optimization via Deep Neural Networks

Shibo Li, Wei Xing, Robert Kirby, Shandian Zhe

Bayesian optimization (BO) is a popular framework for optimizing black-box functions. In many applications,  the objective function can be evaluated at multiple fidelities to enable a trade-off between the cost and accuracy. To reduce the o

ptimization cost, many multi-fidelity BO methods have been proposed. Despite the
ir success, these methods either ignore or over-simplify the strong, complex cor
relations across the fidelities. While the acquisition function is therefore eas
y and convenient to calculate,  these methods can be inefficient in estimating t
he objective function. To address this issue, we propose Deep Neural Network Mul
ti-Fidelity Bayesian Optimization (DNN-MFBO) that can flexibly capture all kinds
 of complicated relationships between the fidelities to improve the objective fu
nction estimation and hence the optimization performance. We use sequential,  fi
delity-wise Gauss-Hermite quadrature and moment-matching to compute a mutual inf
ormation-based acquisition function in a tractable and highly efficient way. We
show the advantages of our method in both synthetic benchmark datasets and real-
world applications in engineering design.
**********

PlanGAN: Model-based Planning With Sparse Rewards and Multiple Goals
Henry Charlesworth, Giovanni Montana
Learning with sparse rewards remains a significant challenge in reinforcement le
arning (RL), especially when the aim is to train a policy capable of achieving m
ultiple different goals. To date, the most successful approaches for dealing wit
h multi-goal, sparse reward environments have been model-free RL algorithms. In
this work we propose PlanGAN, a model-based algorithm specifically designed for
solving multi-goal tasks in environments with sparse rewards. Our method builds
on the fact that any trajectory of experience collected by an agent contains use
ful information about how to achieve the goals observed during that trajectory.
We use this to train an ensemble of conditional generative models (GANs) to gene
rate plausible trajectories that lead the agent from its current state towards a
 specified goal. We then combine these imagined trajectories into a novel planni
ng algorithm in order to achieve the desired goal as efficiently as possible. Th
e performance of PlanGAN has been tested on a number of robotic navigation/manip
ulation tasks in comparison with a range of model-free reinforcement learning ba
selines, including Hindsight Experience Replay. Our studies indicate that  PlanG
AN can achieve comparable performance whilst being around 4-8 times more sample
efficient.
**********

Bad Global Minima Exist and SGD Can Reach Them
Shengchao Liu, Dimitris Papailiopoulos, Dimitris Achlioptas
Several works have aimed to explain why  overparameterized neural networks gener
alize well when trained by Stochastic Gradient Descent (SGD). The consensus expl
anation that has emerged credits the randomized nature of SGD for the bias of th
e training process towards low-complexity models and, thus, for implicit regular
ization. We take a careful look at this explanation in the context of image clas
sification with common deep neural network architectures. We find that if we do
not regularize \emph{explicitly}, then SGD can be easily made to converge to poo
rly-generalizing, high-complexity models: all it takes is to first train on a ra
ndom labeling on the data, before switching to properly training with the correc
t labels. In contrast, we find that in the presence of explicit regularization,
pretraining with random labels has no detrimental effect on SGD. We believe that
 our results give evidence that explicit regularization plays a far more importa
nt role in the success of overparameterized neural networks than what has been u
nderstood until now. Specifically, in suppressing complicated models that got lu
cky with the training data, regularization not only makes simple models that fit
 the data well the global optima, but it also clears the way to make them discov
erable by local methods, such as SGD.
**********

Optimal Prediction of the Number of Unseen Species with Multiplicity
Yi Hao, Ping 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.
**********

Characterizing Optimal Mixed Policies: Where to Intervene and What to Observe
Sanghack Lee, Elias Bareinboim
Intelligent agents are continuously faced with the challenge of optimizing a pol
icy based on what they can observe (see) and which actions they can take (do) in
 the environment where they are deployed. Most policy can be parametrized in ter
ms of these two dimensions, i.e., as a function of what can be seen and done giv
en a certain situation, which we call a \textit{mixed policy}. In this paper, we
 investigate several properties of the class of mixed policies and provide an ef
ficient and effective characterization, including optimality and non-redundancy.
 Specifically, we introduce a graphical criterion to identify unnecessary contex
ts for a set of actions, leading to a natural characterization of non-redundancy
 of mixed policies. We then derive sufficient conditions under which one strateg
y can dominate the other with respect to their maximum achievable expected rewar
ds (optimality). This characterization leads to a fundamental understanding of t
he space of mixed policies and a possible refinement of the agent's strategy so
that it converges to the optimum faster and more robustly. One surprising result
 of the causal characterization is that the agent following a more standard appr
oach --- intervening on all intervenable variables and observing all available c
ontexts --- may be hurting itself, and will never achieve an optimal performance
.
**********

Factor Graph Neural Networks
Zhen Zhang, Fan Wu, Wee Sun Lee
Most of the successful deep neural network architectures are
  structured, often consisting of elements like convolutional neural
  networks and gated recurrent neural networks. Recently, graph neural
  networks (GNNs) have been successfully applied to graph-structured data such
  as point cloud and molecular data. These networks often only consider
  pairwise dependencies, as they operate on a graph structure. We
  generalize the GNN into a factor graph neural network
  (FGNN) providing a simple way to incorporate dependencies among
  multiple variables. We show that FGNN is able to represent
  Max-Product belief propagation, an approximate inference method on
  probabilistic graphical models, providing a theoretical understanding on the c
apabilities of FGNN and related GNNs. Experiments on synthetic and real datasets
 demonstrate the potential of the proposed architecture.
**********

A Closer Look at Accuracy vs. Robustness
Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R. Salakhutdinov, Kamalika
 Chaudhuri
Current methods for training robust networks lead to a drop in test accuracy, wh
ich has led prior works to posit that a robustness-accuracy tradeoff may be inev
itable in deep learning. We take a closer look at this phenomenon and first show
 that real image datasets are actually separated. With this property in mind, we
 then prove that robustness and accuracy should both be achievable for benchmark
 datasets through locally Lipschitz functions, and hence, there should be no inh
erent tradeoff between robustness and accuracy. Through extensive experiments wi
th robustness methods, we argue that the gap between theory and practice arises
from two limitations of current methods: either they fail to impose local Lipsch
itzness or they are insufficiently generalized. We explore combining dropout wit
h robust training methods and obtain better generalization. We conclude that ach
ieving robustness and accuracy in practice may require using methods that impose
 local Lipschitzness and augmenting them with deep learning generalization techn
iques.
**********

Curriculum Learning by Dynamic Instance Hardness
Tianyi Zhou, Shengjie Wang, Jeffrey Bilmes
A good teacher can adjust the curriculum based on students' learning history. By
 analogy, in this paper, we study the dynamics of a deep neural network's (DNN)
performance on individual samples during its learning process. The observed prop

erties allow us to develop an adaptive curriculum that leads to faster learning of more accurate models. We introduce dynamic instance hardness (DIH), the exponential moving average of a sample's instantaneous hardness (e.g., a loss, or a change in outputs) over the training history. A low DIH indicates that a model retains knowledge about a sample over time, and implies a flat loss landscape for that sample. Moreover, for DNNs, we find that a sample's DIH early in training predicts its DIH in later stages. Hence, we can train a model using samples with higher DIH and safely ignore those with lower DIH. This motivates a DIH guided curriculum learning (DIHCL). Compared to existing CL methods: (1) DIH is more stable over time than using only instantaneous hardness, which is noisy due to stochastic training and DNN's non-smoothness; (2) DIHCL is computationally inexpensive since it uses only a byproduct of back-propagation and thus does not require extra inference. On 11 datasets, DIHCL significantly outperforms random mini-batch SGD and recent CL methods in terms of efficiency and final performance.

**********

Spin-Weighted Spherical CNNs

Carlos Esteves, Ameesh Makadia, Kostas Daniilidis

Learning equivariant representations is a promising way to reduce sample and model complexity and improve the generalization performance of deep neural networks. The spherical CNNs are successful examples, producing SO(3)-equivariant representations of spherical inputs. There are two main types of spherical CNNs. The first type lifts the inputs to functions on the rotation group SO(3) and applies convolutions on the group, which are computationally expensive since SO(3) has one extra dimension. The second type applies convolutions directly on the sphere, which are limited to zonal (isotropic) filters, and thus have limited expressivity. In this paper, we present a new type of spherical CNN that allows anisotropic filters in an efficient way, without ever leaving the spherical domain. The key idea is to consider spin-weighted spherical functions, which were introduced in physics in the study of gravitational waves. These are complex-valued functions on the sphere whose phases change upon rotation. We define a convolution between spin-weighted functions and build a CNN based on it. The spin-weighted functions can also be interpreted as spherical vector fields, allowing applications to tasks where the inputs or outputs are vector fields. Experiments show that our method outperforms previous methods on tasks like classification of spherical images, classification of 3D shapes and semantic segmentation of spherical panoramas.

**********

Learning to Execute Programs with Instruction Pointer Attention Graph Neural Networks

David Bieber, Charles Sutton, Hugo Larochelle, Daniel Tarlow

Graph neural networks (GNNs) have emerged as a powerful tool for learning software engineering tasks including code completion, bug finding, and program repair. They benefit from leveraging program structure like control flow graphs, but they are not well-suited to tasks like program execution that require far more sequential reasoning steps than number of GNN propagation steps. Recurrent neural networks (RNNs), on the other hand, are well-suited to long sequential chains of reasoning, but they do not naturally incorporate program structure and generally perform worse on the above tasks. Our aim is to achieve the best of both worlds, and we do so by introducing a novel GNN architecture, the Instruction Pointer Attention Graph Neural Networks (IPA-GNN), which achieves improved systematic generalization on the task of learning to execute programs using control flow graphs. The model arises by considering RNNs operating on program traces with branch decisions as latent variables. The IPA-GNN can be seen either as a continuous relaxation of the RNN model or as a GNN variant more tailored to execution. To test the models, we propose evaluating systematic generalization on learning to execute using control flow graphs, which tests sequential reasoning and use of program structure. More practically, we evaluate these models on the task of learning to execute partial programs, as might arise if using the model as a heuristic function in program synthesis. Results show that the IPA-GNN outperforms a variety of RNN and GNN baselines on both tasks.

**********
AutoPrivacy: Automated Layer-wise Parameter Selection for Secure Neural Network Inference
Qian Lou, Song Bian, Lei Jiang
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Baxter Permutation Process
Masahiro Nakano, Akisato Kimura, Takeshi Yamada, Naonori Ueda
In this paper, a Bayesian nonparametric (BNP) model for Baxter permutations (BPs), termed BP process (BPP) is proposed and applied to relational data analysis. The BPs are a well-studied class of permutations, and it has been demonstrated that there is one-to-one correspondence between BPs and several interesting objects including floorplan partitioning (FP), which constitutes a subset of rectangular partitioning (RP). Accordingly, the BPP can be used as an FP model. We combine the BPP with a multi-dimensional extension of the stick-breaking process called the {\it block-breaking process} to fill the gap between FP and RP, and obtain a stochastic process on arbitrary RPs. Compared with conventional BNP models for arbitrary RPs, the proposed model is simpler and has a high affinity with Bayesian inference.
**********
Characterizing emergent representations in a space of candidate learning rules for deep networks
Yinan Cao, Christopher Summerfield, Andrew Saxe
How are sensory representations learned via experience? Deep learning offers a theoretical toolkit for studying how neural codes emerge under different learning rules. Studies suggesting that representations in deep networks resemble those in biological brains have mostly relied on one specific learning rule: gradient descent, the workhorse behind modern deep learning. However, it remains unclear how robust these emergent representations in deep networks are to this specific choice of learning algorithm. Here we present a continuous two-dimensional space of candidate learning rules, parameterized by levels of top-down feedback and Hebbian learning. We show that this space contains five important candidate learning algorithms as specific points--Gradient Descent, Contrastive Hebbian, quasi-Predictive Coding, Hebbian & Anti-Hebbian. Next, we exhaustively characterize the properties of each rule during learning about hierarchically structured data, and identify zones within this space where deep networks exhibit qualitative signatures of biological learning. We find that while a large set of algorithms achieve zero training error at convergence, only a subset show hallmarks of human semantic development like progressive differentiation and illusory correlations. Further, only a subset adjust intermediate neural representations toward task-relevant representations, indicative of backpropagation-like behavior. Finally, we show that algorithms can dramatically differ in their learned neural representations and dynamics, providing experimentally testable hallmarks of different learning principles. Our findings provide a framework linking diverse neural representational geometries to learning principles which can guide future experiments, and offer evidence about the learning rules likely to be at work in biology.
**********
Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation
Rasool Fakoor, Jonas W. Mueller, Nick Erickson, Pratik Chaudhari, Alexander J. Smola
Automated machine learning (AutoML) can produce complex model ensembles by stacking, bagging, and boosting many individual models like trees, deep networks, and nearest neighbor estimators. While highly accurate, the resulting predictors are large, slow, and opaque as compared to their constituents. To improve the deployment of AutoML on tabular data, we propose FAST-DAD to distill arbitrarily-complex ensemble predictors into individual models like boosted trees, random forests, and deep networks. At the heart of our approach is a data augmentation strat

egy based on Gibbs sampling from a self-attention pseudolikelihood estimator. Ac ross 30 datasets spanning regression and binary/multiclass classification tasks, FAST-DAD distillation produces significantly better individual models than one obtains through standard training on the original data. Our individual distilled models are over 10x faster and more accurate than ensemble predictors produced by AutoML tools like H2O/AutoSklearn.
**********

Adaptive Probing Policies for Shortest Path Routing
Aditya Bhaskara, Sreenivas Gollapudi, Kostas Kollias, Kamesh Munagala
Requests for name changes in the electronic proceedings will be accepted with 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 Heavily-Constrained Learning with Lagrange Multiplier Models
Harikrishna Narasimhan, Andrew Cotter, Yichen Zhou, Serena Wang, Wenshuo Guo
In machine learning applications such as ranking fairness or fairness over inter sectional groups, one often encounters optimization problems with an extremely l arge number of constraints. In particular, with ranking fairness tasks, there ma y even be a variable number of constraints, e.g. one for each query in the train ing set. In these cases, the standard approach of optimizing a Lagrangian while maintaining one Lagrange multiplier per constraint may no longer be practical. O ur proposal is to associate a feature vector with each constraint, and to learn a ``multiplier model'' that maps each such vector to the corresponding Lagrange multiplier. We prove optimality, approximate feasibility and generalization guar antees under assumptions on the flexibility of the multiplier model, and empiric ally demonstrate that our method is effective on real-world case studies.
**********

Faster Randomized Infeasible Interior Point Methods for Tall/Wide Linear Program s
Agniva Chowdhury, Palma London, Haim Avron, Petros Drineas
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Sliding Window Algorithms for k-Clustering Problems
Michele Borassi, Alessandro Epasto, Silvio Lattanzi, Sergei Vassilvitskii, Morte za Zadimoghaddam
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

AdaShare: Learning What To Share For Efficient Deep Multi-Task Learning
Ximeng Sun, Rameswar Panda, Rogerio Feris, Kate Saenko
Multi-task learning is an open and challenging problem in computer vision. The t ypical way of conducting multi-task learning with deep neural networks is either through handcrafted schemes that share all initial layers and branch out at an adhoc point, or through separate task-specific networks with an additional featu re sharing/fusion mechanism. Unlike existing methods, we propose an adaptive sharing approach, calledAdaShare, that decides what to share across which tasks to achieve the best recognition accuracy, while taking resource efficiency into account. Specifically, our main idea is to learn the sharing pattern through a t ask-specific policy that selectively chooses which layers to execute for a given task in the multi-task network. We efficiently optimize the task-specific polic y jointly with the network weights, using standard back-propagation. Experiments on several challenging and diverse benchmark datasets with a variable number of tasks well demonstrate the efficacy of our approach over state-of-the-art metho ds. Project page: https://cs-people.bu.edu/sunxm/AdaShare/project.html

**********
Approximate Cross-Validation for Structured Models
Soumya Ghosh, Will Stephenson, Tin D. Nguyen, Sameer Deshpande, Tamara Broderick
Many modern data analyses benefit from explicitly modeling dependence structure in data -- such as measurements across time or space, ordered words in a sentence, or genes in a genome. A gold standard evaluation technique is structured cross-validation (CV), which leaves out some data subset (such as data within a time interval or data in a geographic region) in each fold. But CV here can be prohibitively slow due to the need to re-run already-expensive learning algorithms many times.
Previous work has shown approximate cross-validation (ACV) methods provide a fast and provably accurate alternative in the setting of empirical risk minimization.
But this existing ACV work is restricted to simpler models by the assumptions that (i) data across CV folds are independent and (ii) an exact initial model fit is available. In structured data analyses, both these assumptions are often untrue.
In the present work, we address (i) by extending ACV to CV schemes with dependence structure between the folds.
To address (ii), we verify -- both theoretically and empirically -- that ACV quality deteriorates smoothly with noise in the initial fit. We demonstrate the accuracy and computational benefits of our proposed methods on a diverse set of real-world applications.
**********
Exemplar VAE: Linking Generative Models, Nearest Neighbor Retrieval, and Data Augmentation
Sajad Norouzi, David J. Fleet, Mohammad Norouzi
We introduce Exemplar VAEs, a family of generative models that bridge the gap between parametric and non-parametric, exemplar based generative models. Exemplar VAE is a variant of VAE with a non-parametric latent prior based on a Parzen window estimator. To sample from it, one first draws a random exemplar from a training set, then stochastically transforms that exemplar into a latent code and a new observation. We propose retrieval augmented training (RAT) as a way to speed up Exemplar VAE training by using approximate nearest neighbor search in the latent space to define a lower bound on log marginal likelihood. To enhance generalization, model parameters are learned using exemplar leave-one-out and subsampling. Experiments demonstrate the effectiveness of Exemplar VAEs on density estimation and representation learning.
Importantly, generative data augmentation using Exemplar VAEs on permutation invariant MNIST and Fashion MNIST reduces classification error from 1.17% to 0.69% and from 8.56% to 8.16%.
**********
Debiased Contrastive Learning
Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, Stefanie Jegelka
A prominent technique for self-supervised representation learning has been to contrast semantically similar and dissimilar pairs of samples. Without access to labels, dissimilar (negative) points are typically taken to be randomly sampled datapoints, implicitly accepting that these points may, in reality, actually have the same label. Perhaps unsurprisingly, we observe that sampling negative examples from truly different labels improves performance, in a synthetic setting where labels are available. Motivated by this observation, we develop a debiased contrastive objective that corrects for the sampling of same-label datapoints, even without knowledge of the true labels. Empirically, the proposed objective consistently outperforms the state-of-the-art for representation learning in vision, language, and reinforcement learning benchmarks. Theoretically, we establish generalization bounds for the downstream classification task.
**********
UCSG-NET- Unsupervised Discovering of Constructive Solid Geometry Tree
Kacper Kania, Maciej Zieba, Tomasz Kajdanowicz

Signed distance field (SDF) is a prominent implicit representation of 3D meshes. Methods that are based on such representation achieved state-of-the-art 3D shape reconstruction quality. However, these methods struggle to reconstruct non-convex shapes. One remedy is to incorporate a constructive solid geometry framework (CSG) that represents a shape as a decomposition into primitives. It allows to embody a 3D shape of high complexity and non-convexity with a simple tree representation of Boolean operations. Nevertheless, existing approaches are supervised and require the entire CSG parse tree that is given upfront during the training process. On the contrary, we propose a model that extracts a CSG parse tree without any supervision - UCSG-Net. Our model predicts parameters of primitives and binarizes their SDF representation through differentiable indicator function. It is achieved jointly with discovering the structure of a Boolean operators tree. The model selects dynamically which operator combination over primitives leads to the reconstruction of high fidelity. We evaluate our method on 2D and 3D autoencoding tasks. We show that the predicted parse tree representation is interpretable and can be used in CAD software.
**********

Generalized Boosting
Arun Suggala, Bingbin Liu, Pradeep Ravikumar
Boosting is a widely used learning technique in machine learning for solving classification problems. In boosting, one predicts the label of an example using an ensemble of weak classifiers. While boosting has shown tremendous success on many classification problems involving tabular data, it performs poorly on complex classification tasks involving low-level features such as image classification tasks. This drawback stems from the fact that boosting builds an additive model of weak classifiers, each of which has very little predictive power. Often, the resulting additive models are not powerful enough to approximate the complex decision boundaries of real-world classification problems. In this work, we present a general framework for boosting where, similar to traditional boosting, we aim to boost the performance of a weak learner and transform it into a strong learner. However, unlike traditional boosting, our framework allows for more complex forms of aggregation of weak learners. In this work, we specifically focus on one form of aggregation - \emph{function composition}. We show that many popular greedy algorithms for learning deep neural networks (DNNs) can be derived from our framework using function compositions for aggregation. Moreover, we identify the drawbacks of these greedy algorithms and propose new algorithms that fix these issues. Using thorough empirical evaluation, we show that our learning algorithms have superior performance over traditional additive boosting algorithms, as well as existing greedy learning techniques for DNNs. An important feature of our algorithms is that they come with strong theoretical guarantees.
**********

COT-GAN: Generating Sequential Data via Causal Optimal Transport
Tianlin Xu, Li Kevin Wenliang, Michael Munn, Beatrice Acciaio
We introduce COT-GAN, an adversarial algorithm to train implicit generative models optimized for producing sequential data. The loss function of this algorithm is formulated using ideas from Causal Optimal Transport (COT), which combines classic optimal transport methods with an additional temporal causality constraint. Remarkably, we find that this causality condition provides a natural framework to parameterize the cost function that is learned by the discriminator as a robust (worst-case) distance, and an ideal mechanism for learning time dependent data distributions. Following Genevay et al. (2018), we also include an entropic penalization term which allows for the use of the Sinkhorn algorithm when computing the optimal transport cost.  Our experiments show effectiveness and stability of COT-GAN when generating both low- and high-dimensional time-series data. The success of the algorithm also relies on a new, improved version of the Sinkhorn divergence which demonstrates less bias in learning.
**********

Impossibility Results for Grammar-Compressed Linear Algebra
Amir Abboud, Arturs Backurs, Karl Bringmann, Marvin Künnemann
To handle vast amounts of data, it is natural and popular to compress vectors an

d matrices.  When we compress a vector from size N down to size n << N, it certainly makes it easier to store and transmit efficiently, but does it also make it easier to process?
**********

Understanding spiking networks through convex optimization
Allan Mancoo, Sander Keemink, Christian K. Machens
Neurons mainly communicate through spikes, and much effort has been spent to understand how the dynamics of spiking neural networks (SNNs) relates to their connectivity. Meanwhile, most major advances in machine learning have been made with simpler, rate-based networks, with SNNs only recently showing competitive results, largely thanks to transferring insights from rate to spiking networks. However, it is still an open question exactly which computations SNNs perform. Recently, the time-averaged firing rates of several SNNs were shown to yield the solutions to convex optimization problems. Here we turn these findings around and show that virtually all inhibition-dominated SNNs can be understood through the lens of convex optimization, with network connectivity, timescales, and firing thresholds being intricately linked to the parameters of underlying convex optimization problems. This approach yields new, geometric insights into the computations performed by spiking networks. In particular, we establish a class of SNNs whose instantaneous output provides a solution to linear or quadratic programming problems, and we thereby reveal their input-output mapping. Using these insights, we derive local, supervised learning rules that can approximate given convex input-output functions, and we show that the resulting networks are consistent with many features from biological networks, such as low firing rates, irregular firing, E/I balance, and robustness to perturbations and synaptic delays.
**********

Better Full-Matrix Regret via Parameter-Free Online Learning
Ashok Cutkosky
We provide online convex optimization algorithms that guarantee improved full-matrix regret bounds. These algorithms extend prior work in several ways. First, we seamlessly allow for the incorporation of constraints without requiring unknown oracle-tuning for any learning rate parameters. Second, we improve the regret of the full-matrix AdaGrad algorithm by suggesting a better learning rate value and showing how to tune the learning rate to this value on-the-fly. Third, all our bounds are obtained via a general framework for constructing regret bounds that depend on an arbitrary sequence of norms.
**********

Large-Scale Methods for Distributionally Robust Optimization
Daniel Levy, Yair Carmon, John C. Duchi, 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.
**********

Analysis and Design of Thompson Sampling for Stochastic Partial Monitoring
Taira Tsuchiya, Junya Honda, Masashi Sugiyama
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Bandit Linear Control
Asaf Cassel, Tomer Koren
We consider the problem of controlling a known linear dynamical system under stochastic noise, adversarially chosen costs, and bandit feedback. Unlike the full feedback setting where the entire cost function is revealed after each decision, here only the cost incurred by the learner is observed. We present a new and efficient algorithm that, for strongly convex and smooth costs, obtains regret that grows with the square root of the time horizon T. We also give extensions of this result to general convex, possibly non-smooth costs, and to non-stochastic s

ystem noise. A key component of our algorithm is a new technique for addressing bandit optimization of loss functions with memory.
**********
Refactoring Policy for Compositional Generalizability using Self-Supervised Object Proposals
Tongzhou Mu, Jiayuan Gu, Zhiwei Jia, Hao Tang, Hao Su
We study how to learn a policy with compositional generalizability. We propose a two-stage framework, which refactorizes a high-reward teacher policy into a generalizable student policy with strong inductive bias. Particularly, we implement an object-centric GNN-based student policy, whose input objects are learned from images through self-supervised learning. Empirically, we evaluate our approach on four difficult tasks that require compositional generalizability, and achieve superior performance compared to baselines.
**********
PEP: Parameter Ensembling by Perturbation
Alireza Mehrtash, Purang Abolmaesumi, Polina Golland, Tina Kapur, Demian Wassermann, William Wells
Ensembling is now recognized as an effective approach for increasing the predictive performance and calibration of deep networks. We introduce a new approach, Parameter Ensembling by Perturbation (PEP), that constructs an ensemble of parameter values as random perturbations of the optimal parameter set from training by a Gaussian with a single variance parameter. The variance is chosen to maximize the log-likelihood of the ensemble average (■) on the validation data set. Empirically, and perhaps surprisingly, ■ has a well-defined maximum as the variance grows from zero (which corresponds to the baseline model). Conveniently, calibration level of predictions also tends to grow favorably until the peak of ■ is reached. In most experiments, PEP provides a small improvement in performance, and, in some cases, a substantial improvement in empirical calibration. We show that this "PEP effect'' (the gain in log-likelihood) is related to the mean curvature of the likelihood function and the empirical Fisher information. Experiments on ImageNet pre-trained networks including ResNet, DenseNet, and Inception showed improved calibration and likelihood. We further observed a mild improvement in classification accuracy on these networks. Experiments on classification benchmarks such as MNIST and CIFAR-10 showed improved calibration and likelihood, as well as the relationship between the PEP effect and overfitting; this demonstrates that PEP can be used to probe the level of overfitting that occurred during training. In general, no special training procedure or network architecture is needed, and in the case of pre-trained networks, no additional training is needed.
**********
Theoretical Insights Into Multiclass Classification: A High-dimensional Asymptotic View
Christos Thrampoulidis, Samet Oymak, Mahdi Soltanolkotabi
Contemporary machine learning applications often involve classification tasks with many classes. Despite their extensive use, a precise understanding of the statistical properties and behavior of classification algorithms is still missing, especially in modern regimes where the number of classes is rather large. In this paper, we take a step in this direction by providing the first asymptotically precise analysis of linear multiclass classification. Our theoretical analysis allows us to precisely characterize how the test error varies over different training algorithms, data distributions, problem dimensions as well as number of classes, inter/intra class correlations and class priors. Specifically, our analysis reveals that the classification accuracy is highly distribution-dependent with different algorithms achieving optimal performance for different data distributions and/or training/features sizes. Unlike linear regression/binary classification, the test error in multiclass classification relies on intricate functions of the trained model (e.g., correlation between some of the trained weights) whose asymptotic behavior is difficult to characterize. This challenge is already present in simple classifiers, such as those minimizing a square loss. Our novel theoretical techniques allow us to overcome some of these challenges. The insight

s gained may pave the way for a precise understanding of other classification al
gorithms beyond those studied in this paper.
**********

Adversarial Example Games
Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon L
acoste-Julien, Will Hamilton
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Residual Distillation: Towards Portable Deep Neural Networks without Shortcuts
Guilin Li, Junlei Zhang, Yunhe Wang, Chuanjian Liu, Matthias Tan, Yunfeng Lin, W
ei Zhang, Jiashi Feng, Tong Zhang
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Provably Efficient Neural Estimation of Structural Equation Models: An Adversari
al Approach
Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Mladen Kolar, Zhaoran Wang
Structural equation models (SEMs) are widely
used in sciences, ranging from economics to psychology,
to uncover causal relationships underlying a complex system
under consideration and estimate structural parameters of interest.
We study estimation in a class of generalized SEMs where the object
of interest is defined as the solution to a linear operator equation.
We formulate the linear operator equation as a min-max game, where both
players are parameterized by neural networks (NNs), and learn the
parameters of these neural networks using the stochastic gradient descent.
We consider both 2-layer and multi-layer NNs with ReLU activation
functions and prove global convergence in an overparametrized regime, where
the number of neurons is diverging. The results are established using
techniques from online learning and local linearization of NNs,
and improve in several aspects the current state-of-the-art. For the first
time we provide a tractable estimation procedure for SEMs
based on NNs with provable convergence and without the need for sample
splitting.
**********

Security Analysis of Safe and Seldonian Reinforcement Learning Algorithms
Pinar Ozisik, Philip S. Thomas
We analyze the extent to which existing methods rely on accurate training data f
or a specific class of reinforcement learning (RL) algorithms, known as Safe and
 Seldonian RL. We introduce a new measure of security to quantify the susceptibi
lity to perturbations in training data by creating an attacker model that repres
ents a worst-case analysis, and show that a couple of Seldonian RL methods are e
xtremely sensitive to even a few data corruptions. We then introduce a new algor
ithm that is more robust against data corruptions, and demonstrate its usage in
practice on some RL problems, including a grid-world and a diabetes treatment si
mulation.
**********

Learning to Play Sequential Games versus Unknown Opponents
Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, Andreas Krause
We consider a repeated sequential game between a learner, who plays first, and a
n opponent who responds to the chosen action. We seek to design strategies for t
he learner to successfully interact with the opponent. While most previous appro
aches consider known opponent models, we focus on the setting in which the oppon
ent's model is unknown. To this end, we use kernel-based regularity assumptions
to capture and exploit the structure in the opponent's response. We propose a no

vel algorithm for the learner when playing against an adversarial sequence of opponents. The algorithm combines ideas from bilevel optimization and online learning to effectively balance between exploration (learning about the opponent's model) and exploitation (selecting highly rewarding actions for the learner). Our results include algorithm's regret guarantees that depend on the regularity of the opponent's response and scale sublinearly with the number of game rounds. Moreover, we specialize our approach to repeated Stackelberg games, and empirically demonstrate its effectiveness in a traffic routing and wildlife conservation task.

**********

Further Analysis of Outlier Detection with Deep Generative Models

Ziyu Wang, Bin Dai, David Wipf, Jun Zhu

The recent, counter-intuitive discovery that deep generative models (DGMs) can frequently assign a higher likelihood to outliers has implications for both outlier detection applications as well as our overall understanding of generative modeling. In this work, we present a possible explanation for this phenomenon, starting from the observation that a model's typical set and high-density region may not conincide. From this vantage point we propose a novel outlier test, the empirical success of which suggests that the failure of existing likelihood-based outlier tests does not necessarily imply that the corresponding generative model is uncalibrated. We also conduct additional experiments to help disentangle the impact of low-level texture versus high-level semantics in differentiating outliers. In aggregate, these results suggest that modifications to the standard evaluation practices and benchmarks commonly applied in the literature are needed.

**********

Bridging Imagination and Reality for Model-Based Deep Reinforcement Learning

Guangxiang Zhu, Minghao Zhang, Honglak Lee, Chongjie Zhang

Sample efficiency has been one of the major challenges for deep reinforcement learning. Recently, model-based reinforcement learning has been proposed to address this challenge by performing planning on imaginary trajectories with a learned world model. However, world model learning may suffer from overfitting to training trajectories, and thus model-based value estimation and policy search will be prone to be sucked in an inferior local policy. In this paper, we propose a novel model-based reinforcement learning algorithm, called BrIdging Reality and Dream (BIRD). It maximizes the mutual information between imaginary and real trajectories so that the policy improvement learned from imaginary trajectories can be easily generalized to real trajectories. We demonstrate that our approach improves sample efficiency of model-based planning, and achieves state-of-the-art performance on challenging visual control benchmarks.

**********

Neural Networks Learning and Memorization with (almost) no Over-Parameterization

Amit Daniely

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

**********

Exploiting Higher Order Smoothness in Derivative-free Optimization and Continuous Bandits

Arya Akhavan, Massimiliano Pontil, Alexandre Tsybakov

We address the problem of zero-order optimization of a strongly convex function. The goal is to find the minimizer of the function by a sequential exploration of its function values, under measurement noise. We study the impact of higher order smoothness properties of the function on the optimization error and on the online regret. To solve this problem we consider a randomized approximation of the projected gradient descent algorithm. The gradient is estimated by a randomized procedure involving two function evaluations and a smoothing kernel. We derive upper bounds for this algorithm both in the constrained and unconstrained settings and prove minimax lower bounds for any sequential search method. Our results

imply that the zero-order algorithm is nearly optimal in terms of sample complexity and the problem parameters. Based on this algorithm, we also propose an estimator of the minimum value of the function achieving almost sharp oracle behavior. We compare our results with the state-of-the-art, highlighting a number of key improvements.

**********

Towards a Combinatorial Characterization of Bounded-Memory Learning

Alon Gonen, Shachar Lovett, Michal Moshkovitz

Combinatorial dimensions play an important role in the theory of machine learning. For example, VC dimension characterizes PAC learning, SQ dimension characterizes weak learning with statistical queries, and Littlestone dimension characterizes online learning. In this paper we aim to develop combinatorial dimensions that characterize bounded memory learning. We propose a candidate solution for the case of realizable strong learning under a known distribution, based on the SQ dimension of neighboring distributions. We prove both upper and lower bounds for our candidate solution, that match in some regime of parameters. This is the first characterization of strong learning under space constraints in any regime. In this parameter regime there is an equivalence between bounded memory and SQ learning. We conjecture that our characterization holds in a much wider regime of parameters.

**********

Chaos, Extremism and Optimism: Volume Analysis of Learning in Games

Yun Kuen Cheung, Georgios Piliouras

We perform volume analysis of Multiplicative Weights Updates (MWU) and its optimistic variant (OMWU) in zero-sum as well as coordination games. Our analysis provides new insights into these game/dynamical systems, which seem hard to achieve via the classical techniques within Computer Science and Machine Learning.

**********

On Regret with Multiple Best Arms

Yinglun Zhu, Robert Nowak

We study a regret minimization problem with the existence of multiple best/near-optimal arms in the multi-armed bandit setting. We consider the case when the number of arms/actions is comparable or much larger than the time horizon, and make no assumptions about the structure of the bandit instance. Our goal is to design algorithms that can automatically adapt to the unknown hardness of the problem, i.e., the number of best arms. Our setting captures many modern applications of bandit algorithms where the action space is enormous and the information about the underlying instance/structure is unavailable. We first propose an adaptive algorithm that is agnostic to the hardness level and theoretically derive its regret bound. We then prove a lower bound for our problem setting, which indicates: (1) no algorithm can be minimax optimal simultaneously over all hardness levels; and (2) our algorithm achieves a rate function that is Pareto optimal. With additional knowledge of the expected reward of the best arm, we propose another adaptive algorithm that is minimax optimal, up to polylog factors, over all hardness levels. Experimental results confirm our theoretical guarantees and show advantages of our algorithms over the previous state-of-the-art.

**********

Matrix Completion with Hierarchical Graph Side Information

Adel Elmahdy, Junhyung Ahn, Changho Suh, Soheil Mohajer

We consider a matrix completion problem that exploits social or item similarity graphs as side information. We develop a universal, parameter-free, and computationally efficient algorithm that starts with hierarchical graph clustering and then iteratively refines estimates both on graph clustering and matrix ratings. Under a hierarchical stochastic block model that well respects practically-relevant social graphs and a low-rank rating matrix model (to be detailed), we demonstrate that our algorithm achieves the information-theoretic limit on the number of observed matrix entries (i.e., optimal sample complexity) that is derived by maximum likelihood estimation together with a lower-bound impossibility result. One consequence of this result is that exploiting the hierarchical structure of social graphs yields a substantial gain in sample complexity relative to the one

that simply identifies different groups without resorting to the relational stru
cture across them. We conduct extensive experiments both on synthetic and real-w
orld datasets to corroborate our theoretical results as well as to demonstrate s
ignificant performance improvements over other matrix completion algorithms that
 leverage graph side information.
**********

Is Long Horizon RL More Difficult Than Short Horizon RL?
Ruosong Wang, Simon S. Du, Lin Yang, Sham Kakade
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: B
ayesian inference for latent Gaussian models and beyond
Charles Margossian, Aki Vehtari, Daniel Simpson, Raj Agrawal
Gaussian latent variable models are a key class of Bayesian hierarchical models
with applications in many fields. Performing Bayesian inference on such models c
an be challenging as Markov chain Monte Carlo algorithms struggle with the geome
try of the resulting posterior distribution and can be prohibitively slow. An al
ternative is to use a Laplace approximation to marginalize out the latent Gaussi
an variables and then integrate out the remaining hyperparameters using dynamic
Hamiltonian Monte Carlo, a gradient-based Markov chain Monte Carlo sampler. To i
mplement this scheme efficiently, we derive a novel adjoint method that propagat
es the minimal information needed to construct the gradient of the approximate m
arginal likelihood. This strategy yields a scalable differentiation method that
is orders of magnitude faster than state of the art differentiation techniques w
hen the hyperparameters are high dimensional. We prototype the method in the pro
babilistic programming framework Stan and test the utility of the embedded Lapla
ce approximation on several models, including one where the dimension of the hyp
erparameter is ~6,000. Depending on the cases, the benefits can include an allev
iation of the geometric pathologies that frustrate Hamiltonian Monte Carlo and a
 dramatic speed-up.
**********

Adversarial Learning for Robust Deep Clustering
Xu Yang, Cheng Deng, Kun Wei, Junchi Yan, Wei Liu
Deep clustering integrates embedding and clustering together to obtain the optim
al nonlinear embedding space, which is more effective in real-world scenarios co
mpared with conventional clustering methods. However, the robustness of the clus
tering network is prone to being attenuated especially when it encounters an adv
ersarial attack. A small perturbation in the embedding space will lead to divers
e clustering results since the labels are absent. In this paper, we propose a ro
bust deep clustering method based on adversarial learning. Specifically, we firs
t attempt to define adversarial samples in the embedding space for the clusterin
g network. Meanwhile, we devise an adversarial attack strategy to explore sample
s that easily fool the clustering layers but do not impact the performance of th
e deep embedding. We then provide a simple yet efficient defense algorithm to im
prove the robustness of the clustering network. Experimental results on two popu
lar datasets show that the proposed adversarial learning method can significantl
y enhance the robustness and further improve the overall clustering performance.
 Particularly, the proposed method is generally applicable to multiple existing
clustering frameworks to boost their robustness. The source code is available at
 https://github.com/xdxuyang/ALRDC.
**********

Learning Mutational Semantics
Brian Hie, Ellen Zhong, Bryan Bryson, Bonnie Berger
In many natural domains, changing a small part of an entity can transform its se
mantics; for example, a single word change can alter the meaning of a sentence,
or a single amino acid change can mutate a viral protein to escape antiviral tre
atment or immunity. Although identifying such mutations can be desirable (for ex

ample, therapeutic design that anticipates avenues of viral escape), the rules g
overning semantic change are often hard to quantify. Here, we introduce the prob
lem of identifying mutations with a large effect on semantics, but where valid m
utations are under complex constraints (for example, English grammar or biologic
al viability), which we refer to as constrained semantic change search (CSCS). W
e propose an unsupervised solution based on language models that simultaneously
learn continuous latent representations. We report good empirical performance on
 CSCS of single-word mutations to news headlines, map a continuous semantic spac
e of viral variation, and, notably, show unprecedented zero-shot prediction of s
ingle-residue escape mutations to key influenza and HIV proteins, suggesting a p
roductive link between modeling natural language and pathogenic evolution.
**********

Learning to Learn Variational Semantic Memory
Xiantong Zhen, Yingjun Du, Huan Xiong, Qiang Qiu, Cees Snoek, Ling Shao
In this paper, we introduce variational semantic memory into meta-learning to ac
quire long-term knowledge for few-shot learning. The variational semantic memory
 accrues and stores semantic information for the probabilistic inference of clas
s prototypes in a hierarchical Bayesian framework. The semantic memory is grown
from scratch and gradually consolidated by absorbing information from tasks it e
xperiences. By doing so, it is able to accumulate long-term, general knowledge t
hat enables it to learn new concepts of objects.
We formulate memory recall as the variational inference of a latent memory varia
ble from addressed contents, which offers a principled way to adapt the knowledg
e to individual tasks. Our variational semantic memory, as a new long-term memor
y module, confers principled recall and update mechanisms that enable semantic i
nformation to be efficiently accrued and adapted for few-shot learning.
Experiments demonstrate that the probabilistic modelling of prototypes achieves
a more informative representation of object classes compared to deterministic ve
ctors. The consistent new state-of-the-art performance on four benchmarks shows
the benefit of variational semantic memory in boosting few-shot recognition.
**********

Myersonian Regression
Allen Liu, Renato Leme, Jon Schneider
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Learnability with Indirect Supervision Signals
Kaifu Wang, Qiang Ning, Dan Roth
Learning from indirect supervision signals is important in real-world AI applica
tions when, often, gold labels are missing or too costly. In this paper, we deve
lop a unified theoretical framework for multi-class classification when the supe
rvision is provided by a variable that contains nonzero mutual information with
the gold label. The nature of this problem is determined by (i) the transition p
robability from the gold labels to the indirect supervision variables and (ii) t
he learner's prior knowledge about the transition. Our framework relaxes assumpt
ions made in the literature, and supports learning with unknown, non-invertible
and instance-dependent transitions. Our theory introduces a novel concept called
 \emph{separation}, which characterizes the learnability and generalization boun
ds. We also demonstrate the application of our framework via concrete novel resu
lts in a variety of learning scenarios such as learning with superset annotation
s and joint supervision signals.
**********

Towards Safe Policy Improvement for Non-Stationary MDPs
Yash Chandak, Scott Jordan, Georgios Theocharous, Martha White, Philip S. Thomas
Many real-world sequential decision-making problems involve critical systems wit
h financial risks and human-life risks. While several works in the past have pro
posed methods that are safe for deployment, they assume that the underlying prob
lem is stationary. However, many real-world problems of interest exhibit non-sta

tionarity, and when stakes are high, the cost associated with a false stationarity assumption may be unacceptable. We take the first steps towards ensuring safety, with high confidence, for smoothly-varying non-stationary decision problems. Our proposed method extends a type of safe algorithm, called a Seldonian algorithm, through a synthesis of model-free reinforcement learning with time-series analysis. Safety is ensured using sequential hypothesis testing of a policy's forecasted performance, and confidence intervals are obtained using wild bootstrap.
**********

Finer Metagenomic Reconstruction via Biodiversity Optimization
Simon Foucart, David Koslicki
When analyzing communities of microorganisms from their sequenced DNA, an important task is taxonomic profiling: enumerating the presence and relative abundance of all organisms, or merely of all taxa, contained in the sample. This task can be tackled via compressive-sensing-based approaches, which favor communities featuring the fewest organisms among those consistent with the observed DNA data. Despite their successes, these parsimonious approaches sometimes conflict with biological realism by overlooking organism similarities. Here, we leverage a recently developed notion of biological diversity that simultaneously accounts for organism similarities and retains the optimization strategy underlying compressive-sensing-based approaches. We demonstrate that minimizing biological diversity still produces sparse taxonomic profiles and we experimentally validate superiority to existing compressive-sensing-based approaches. Despite showing that the objective function is almost never convex and often concave, generally yielding NP-hard problems, we exhibit ways of representing organism similarities for which minimizing diversity can be performed via a sequence of linear programs guaranteed to decrease diversity. Better yet, when biological similarity is quantified by k-mer co-occurrence (a popular notion in bioinformatics), minimizing diversity actually reduces to one linear program that can utilize multiple k-mer sizes to enhance performance. In proof-of-concept experiments, we verify that the latter procedure can lead to significant gains when taxonomically profiling a metagenomic sample, both in terms of reconstruction accuracy and computational performance.
**********

Causal Discovery in Physical Systems from Videos
Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, Animesh Garg
Causal discovery is at the core of human cognition. It enables us to reason about the environment and make counterfactual predictions about unseen scenarios that can vastly differ from our previous experiences. We consider the task of causal discovery from videos in an end-to-end fashion without supervision on the ground-truth graph structure. In particular, our goal is to discover the structural dependencies among environmental and object variables: inferring the type and strength of interactions that have a causal effect on the behavior of the dynamical system. Our model consists of (a) a perception module that extracts a semantically meaningful and temporally consistent keypoint representation from images, (b) an inference module for determining the graph distribution induced by the detected keypoints, and (c) a dynamics module that can predict the future by conditioning on the inferred graph. We assume access to different configurations and environmental conditions, i.e., data from unknown interventions on the underlying system; thus, we can hope to discover the correct underlying causal graph without explicit interventions. We evaluate our method in a planar multi-body interaction environment and scenarios involving fabrics of different shapes like shirts and pants. Experiments demonstrate that our model can correctly identify the interactions from a short sequence of images and make long-term future predictions. The causal structure assumed by the model also allows it to make counterfactual predictions and extrapolate to systems of unseen interaction graphs or graphs of various sizes.
**********

Glyph: Fast and Accurately Training Deep Neural Networks on Encrypted Data
Qian Lou, Bo Feng, Geoffrey  Charles Fox, Lei Jiang
Because of the lack of expertise, to gain benefits from their data, average user

s have to upload their private data to cloud servers they may not trust. Due to legal or privacy constraints, most users are willing to contribute only their encrypted data, and lack interests or resources to join deep neural network (DNN) training in cloud. To train a DNN on encrypted data in a completely non-interactive way, a recent work proposes a fully homomorphic encryption (FHE)-based technique implementing all activations by \textit{Brakerski-Gentry-Vaikuntanathan} (BGV)-based lookup tables. However, such inefficient lookup-table-based activations significantly prolong private training latency of DNNs.
**********
Smoothed Analysis of Online and Differentially Private Learning
Nika Haghtalab, Tim Roughgarden, Abhishek Shetty
Practical and pervasive needs for robustness and privacy in algorithms have inspired the design of online adversarial and differentially private learning algorithms. The primary quantity that characterizes learnability in these settings is the Littlestone dimension of the class of hypotheses [Ben-David et al., 2009, Alon et al., 2019]. This characterization is often interpreted as an impossibility result because classes such as linear thresholds and neural networks have infinite Littlestone dimension. In this paper, we apply the framework of smoothed analysis [Spielman and Teng, 2004], in which adversarially chosen inputs are perturbed slightly by nature. We show that fundamentally stronger regret and error guarantees are possible with smoothed adversaries than with worst-case adversaries. In particular, we obtain regret and privacy error bounds that depend only on the VC dimension and the bracketing number of a  hypothesis class, and on the magnitudes of the perturbations.
**********
Self-Paced Deep Reinforcement Learning
Pascal Klink, Carlo D'Eramo, Jan R. Peters, Joni Pajarinen
Curriculum reinforcement learning (CRL) improves the learning speed and stability of an agent by exposing it to a tailored series of tasks throughout learning. Despite empirical successes, an open question in CRL is how to automatically generate a curriculum for a given reinforcement learning (RL) agent, avoiding manual design. In this paper, we propose an answer by interpreting the curriculum generation as an inference problem, where distributions over tasks are progressively learned to approach the target task. This approach leads to an automatic curriculum generation, whose pace is controlled by the agent, with solid theoretical motivation and easily integrated with deep RL algorithms. In the conducted experiments, the curricula generated with the proposed algorithm significantly improve learning performance across several environments and deep RL algorithms, matching or outperforming state-of-the-art existing CRL algorithms.
**********
Kalman Filtering Attention for User Behavior Modeling in CTR Prediction
Hu Liu, Jing LU, Xiwei Zhao, Sulong Xu, Hao Peng, Yutong Liu, Zehua Zhang, Jian Li, Junsheng Jin, Yongjun Bao, Weipeng Yan
Click-through rate (CTR) prediction is one of the fundamental tasks for e-commerce search engines. As search becomes more personalized, it is necessary to capture the user interest from rich behavior data. Existing user behavior modeling algorithms develop different attention mechanisms to emphasize query-relevant behaviors and suppress irrelevant ones. Despite being extensively studied, these attentions still suffer from two limitations. First, conventional attentions mostly limit the attention field only to a single user's behaviors, which is not suitable in e-commerce where users often hunt for new demands that are irrelevant to any historical behaviors. Second, these attentions are usually biased towards frequent behaviors, which is unreasonable since high frequency does not necessarily indicate great importance. To tackle the two limitations, we propose a novel attention mechanism, termed Kalman Filtering Attention (KFAtt), that considers the weighted pooling in attention as a maximum a posteriori (MAP) estimation. By incorporating a priori, KFAtt resorts to global statistics when few user behaviors are relevant. Moreover, a frequency capping mechanism is incorporated to correct the bias towards frequent behaviors. Offline experiments on both benchmark and a 10 billion scale real production dataset, together with an Online A/B test,

show that KFAtt outperforms all compared state-of-the-arts. KFAtt has been deplo
yed in the ranking system of JD.com, one of the largest B2C e-commerce websites
in China, serving the main traffic of hundreds of millions of active users.
**********

Towards Maximizing the Representation Gap between In-Domain & Out-of-Distributio
n Examples

Jay Nandy, Wynne Hsu, Mong Li Lee

Among existing uncertainty estimation approaches, Dirichlet Prior Network (DPN)
distinctly models different predictive uncertainty types.
However, for in-domain examples with high data uncertainties among multiple clas
ses, even a DPN model often produces indistinguishable representations from the
out-of-distribution (OOD) examples, compromising their OOD detection performance
. We address this shortcoming by proposing a novel loss function for DPN to maxi
mize the representation gap between in-domain and OOD examples. Experimental res
ults demonstrate that our proposed approach consistently improves OOD detection
performance.
**********

Fully Convolutional Mesh Autoencoder using Efficient Spatially Varying Kernels

Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, Yaser
 Sheikh

Learning latent representations of registered meshes is useful for many 3D tasks
. Techniques have recently shifted to neural mesh autoencoders. Although they de
monstrate higher precision than traditional methods, they remain unable to captu
re fine-grained deformations. Furthermore, these methods can only be applied to
a template-specific surface mesh, and is not applicable to more general meshes,
like tetrahedrons and non-manifold meshes. While more general graph convolution
methods can be employed, they lack performance in reconstruction precision and r
equire higher memory usage.
In this paper, we propose a non-template-specific fully convolutional mesh autoe
ncoder for arbitrary registered mesh data. It is enabled by our novel convolutio
n and (un)pooling operators learned with globally shared weights and locally var
ying coefficients which can efficiently capture the spatially varying contents p
resented by irregular mesh connections.
Our model outperforms state-of-the-art methods on reconstruction accuracy. In ad
dition, the latent codes of our network are fully localized thanks to the fully
convolutional structure, and thus have much higher interpolation capability than
 many traditional 3D mesh generation models.
**********

GNNGuard: Defending Graph Neural Networks against Adversarial Attacks

Xiang Zhang, Marinka Zitnik

Deep learning methods for graphs achieve remarkable performance on many tasks. H
owever, despite the proliferation of such methods and their success, recent find
ings indicate that small, unnoticeable perturbations of graph structure can cata
strophically reduce performance of even the strongest and most popular Graph Neu
ral Networks (GNNs). Here, we develop GNNGuard, a general defense approach again
st a variety of training-time attacks that perturb the discrete graph structure.
 GNNGuard can be straightforwardly incorporated into any GNN. Its core principle
 is to detect and quantify the relationship between the graph structure and node
 features, if one exists, and then exploit that relationship to mitigate the neg
ative effects of the attack. GNNGuard learns how to best assign higher weights t
o edges connecting similar nodes while pruning edges between unrelated nodes. Th
e revised edges then allow the underlying GNN to robustly propagate neural messa
ges in the graph. GNNGuard introduces two novel components, the neighbor importa
nce estimation, and the layer-wise graph memory, and we show empirically that bo
th components are necessary for a successful defense. Across five GNNs, three de
fense methods, and four datasets, including a challenging human disease graph, e
xperiments show that GNNGuard outperforms existing defense approaches by 15.3% o
n average. Remarkably, GNNGuard can effectively restore state-of-the-art perform
ance of GNNs in the face of various adversarial attacks, including targeted and
non-targeted attacks, and can defend against attacks on heterophily graphs.

**********
Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Single-view Human Reconstruction
Tong He, John Collomosse, Hailin Jin, Stefano Soatto
We propose Geo-PIFu, a method to recover a 3D mesh from a monocular color image of a clothed person. Our method is based on a deep implicit function-based representation to learn latent voxel features using a structure-aware 3D U-Net, to constrain the model in two ways: first, to resolve feature ambiguities in query point encoding, second, to serve as a coarse human shape proxy to regularize the high-resolution mesh and encourage global shape regularity. We show that, by both encoding query points and constraining global shape using latent voxel features, the reconstruction we obtain for clothed human meshes exhibits less shape distortion and improved surface details compared to competing methods. We evaluate Geo-PIFu on a recent human mesh public dataset that is 10x larger than the private commercial dataset used in PIFu and previous derivative work. On average, we exceed the state of the art by 42.7% reduction in Chamfer and Point-to-Surface Distances, and 19.4% reduction in normal estimation errors.
**********
Optimal visual search based on a model of target detectability in natural images
Shima Rashidi, Krista Ehinger, Andrew Turpin, Lars Kulik
To analyse visual systems, the concept of an ideal observer promises an optimal response for a given task. Bayesian ideal observers can provide optimal responses under uncertainty, if they are given the true distributions as input. In visual search tasks, prior studies have used signal to noise ratio (SNR) or psychophysics experiments to set the distributional parameters for simple targets on backgrounds with known patterns, however these methods do not easily translate to complex targets on natural scenes. Here, we develop a model of target detectability in natural images to estimate the parameters of target-present and target-absent distributions for a visual search task. We present a novel approach for approximating the foveated detectability of a known target in natural backgrounds based on biological aspects of human visual system. Our model considers both the uncertainty about target position and the visual system's variability due to its reduced performance in the periphery compared to the fovea. Our automated prediction algorithm uses trained logistic regression as a post processing phase of a pre-trained deep neural network. Eye tracking data from 12 observers detecting targets on natural image backgrounds are used as ground truth to tune foveation parameters and evaluate the model, using cross-validation. Finally, the model of target detectability is used in a Bayesian ideal observer model of visual search, and compared to human search performance.
**********
Towards Convergence Rate Analysis of Random Forests for Classification
Wei Gao, Zhi-Hua Zhou
Random forests have been one of the successful ensemble algorithms in machine learning. The basic idea is to construct a large number of random trees individually and make prediction based on an average of their predictions. The great successes have attracted much attention on the consistency of random forests, mostly focusing on regression. This work takes one step towards convergence rates of random forests for classification. We present the first finite-sample rate O(n^{-1/(8d+2)}) on the convergence of pure random forests for classification, which can be improved to be of O(n^{-1/(3.87d+2)}) by considering the midpoint splitting mechanism. We introduce another variant of random forests, which follow Breiman's original random forests but with different mechanisms on splitting dimensions and positions. We get a convergence rate O(n^{-{1}/(d+2)}(\ln n)^{{1}/(d+2)}) for the variant of random forests, which reaches the minimax rate, except for a factor (\ln n)^{{1}/(d+2)}, of the optimal plug-in classifier under the L-Lipschitz assumption. We achieve tighter convergence rate O(\sqrt{\ln n/n}) under proper assumptions over structural data.
**********
List-Decodable Mean Estimation via Iterative Multi-Filtering
Ilias Diakonikolas, Daniel Kane, Daniel Kongsgaard

Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Recovery of Mangled Clusters with Same-Cluster Queries
Marco Bressan, Nicolò Cesa-Bianchi, Silvio Lattanzi, Andrea Paudice
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Steady State Analysis of Episodic Reinforcement Learning
Huang Bojun
Reinforcement Learning (RL) tasks generally divide into two kinds: continual learning and episodic learning. The concept of steady state has played a foundational role in the continual setting, where unique steady-state distribution is typically presumed to exist in the task being studied, which enables principled conceptual framework as well as efficient data collection method for continual RL algorithms. On the other hand, the concept of steady state has been widely considered irrelevant for episodic RL tasks, in which the decision process terminates in finite time. Alternative concepts, such as episode-wise visitation frequency, are used in episodic RL algorithms, which are not only inconsistent with their counterparts in continual RL, and also make it harder to design and analyze RL algorithms in the episodic setting.
**********

Direct Feedback Alignment Scales to Modern Deep Learning Tasks and Architectures
Julien Launay, Iacopo Poli, François Boniface, Florent Krzakala
Despite being the workhorse of deep learning, the backpropagation algorithm is no panacea. It enforces sequential layer updates, thus preventing efficient parallelization of the training process. Furthermore, its biological plausibility is being challenged. Alternative schemes have been devised; yet, under the constraint of synaptic asymmetry, none have scaled to modern deep learning tasks and architectures. Here, we challenge this perspective, and study the applicability of Direct Feedback Alignment (DFA) to neural view synthesis, recommender systems, geometric learning, and natural language processing. In contrast with previous studies limited to computer vision tasks, our findings show that it successfully trains a large range of state-of-the-art deep learning architectures, with performance close to fine-tuned backpropagation. When a larger gap between DFA and backpropagation exists, like in Transformers, we attribute this to a need to rethink common practices for large and complex architectures. At variance with common beliefs, our work supports that challenging tasks can be tackled in the absence of weight transport.
**********

Bayesian Optimization for Iterative Learning
Vu Nguyen, Sebastian Schulze, Michael Osborne
The performance of deep (reinforcement) learning systems crucially depends on the choice of hyperparameters. Their tuning is notoriously expensive, typically requiring an iterative training process to run for numerous steps to convergence. Traditional tuning algorithms only consider the final performance of hyperparameters acquired after many expensive iterations and ignore intermediate information from earlier training steps. In this paper, we present a Bayesian optimization (BO) approach which exploits the iterative structure of learning algorithms for efficient hyperparameter tuning. We propose to learn an evaluation function compressing learning progress at any stage of the training process into a single numeric score according to both training success and stability. Our BO framework is then trade-off the benefit of assessing a hyperparameter setting over additional training steps against their computation cost. We further increase model efficiency by selectively including scores from different training steps for any evaluated hyperparameter set. We demonstrate the efficiency of our algorithm by tuni

ng hyperparameters for the training of deep reinforcement learning agents and co
nvolutional neural networks. Our algorithm outperforms all existing baselines in
 identifying optimal hyperparameters in minimal time.
**********

Minimax Bounds for Generalized Linear Models
Kuan-Yun Lee, Thomas Courtade
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Projection Robust Wasserstein Distance and Riemannian Optimization
Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, Michael Jordan
Projection robust Wasserstein (PRW) distance, or Wasserstein projection pursuit
(WPP), is a robust variant of the Wasserstein distance. Recent work suggests tha
t this quantity is more robust than the standard Wasserstein distance, in partic
ular when comparing probability measures in high-dimensions. However, it is rule
d out for practical application because the optimization model is essentially no
n-convex and non-smooth which makes the computation intractable. Our contributio
n in this paper is to revisit the original motivation behind WPP/PRW, but take t
he hard route of showing that, despite its non-convexity and lack of nonsmoothne
ss, and even despite some hardness results proved by~\citet{Niles-2019-Estimatio
n} in a minimax sense, the original formulation for PRW/WPP \textit{can} be effi
ciently computed in practice using Riemannian optimization, yielding in relevant
 cases better behavior than its convex relaxation. More specifically, we provide
 three simple algorithms with solid theoretical guarantee on their complexity bo
und (one in the appendix), and demonstrate their effectiveness and efficiency by
 conducing extensive experiments on synthetic and real data. This paper provides
 a first step into a computational theory of the PRW distance and provides the l
inks between optimal transport and Riemannian optimization.
**********

CoinDICE: Off-Policy Confidence Interval Estimation
Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvari, Dale Schuurmans
We study high-confidence behavior-agnostic off-policy evaluation in reinforcemen
t learning, where the goal is to estimate a confidence interval on a target poli
cy's value, given only access to a static experience dataset collected by unknow
n behavior policies. Starting from a function space embedding of the linear prog
ram formulation of the Q-function, we obtain an optimization problem with genera
lized estimating equation constraints. By applying the generalized empirical lik
elihood method to the resulting Lagrangian, we propose CoinDICE, a novel and eff
icient algorithm for computing confidence intervals. Theoretically, we prove the
 obtained confidence intervals are valid, in both asymptotic and finite-sample r
egimes. Empirically, we show in a variety of benchmarks that the confidence inte
rval estimates are tighter and more accurate than existing methods.
**********

Simple and Fast Algorithm for Binary Integer and Online Linear Programming
Xiaocheng Li, Chunlin Sun, Yinyu  Ye
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Learning Diverse and Discriminative Representations via the Principle of Maximal
 Coding Rate Reduction
Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, Yi 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.
**********

Learning Rich Rankings
Arjun Seshadri, Stephen Ragain, Johan Ugander
Although the foundations of ranking are well established, the ranking literature has primarily been focused on simple, unimodal models, e.g. the Mallows and Plackett-Luce models, that define distributions centered around a single total ordering. Explicit mixture models have provided some tools for modelling multimodal ranking data, though learning such models from data is often difficult. In this work, we contribute a contextual repeated selection (CRS) model that leverages recent advances in choice modeling to bring a natural multimodality and richness to the rankings space. We provide rigorous theoretical guarantees for maximum likelihood estimation under the model through structure-dependent tail risk and expected risk bounds. As a by-product, we also furnish the first tight bounds on the expected risk of maximum likelihood estimators for the multinomial logit (MNL) choice model and the Plackett-Luce (PL) ranking model, as well as the first tail risk bound on the PL ranking model. The CRS model significantly outperforms existing methods for modeling real world ranking data in a variety of settings, from racing to rank choice voting.
**********

Color Visual Illusions: A Statistics-based Computational Model
Elad Hirsch, Ayellet Tal
Visual illusions may be explained by the likelihood of patches in real-world images, as argued by input-driven paradigms in Neuro-Science. However, neither the data nor the tools existed in the past to extensively support these explanations. The era of big data opens a new opportunity to study input-driven approaches. We introduce a tool that computes the likelihood of patches, given a large dataset to learn from. Given this tool, we present a model that supports the approach and explains lightness and color visual illusions in a unified manner. Furthermore, our model generates visual illusions in natural images, by applying the same tool, reversely.
**********

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela
Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory can overcome this issue, but have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) -- models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, the other can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state-of-the-art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.
**********

Universal guarantees for decision tree induction via a higher-order splitting criterion
Guy Blanc, Neha Gupta, Jane Lange, Li-Yang Tan
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Trade-offs and Guarantees of Adversarial Representation Learning for Information Obfuscation
Han Zhao, Jianfeng Chi, Yuan Tian, Geoffrey J. Gordon
Crowdsourced data used in machine learning services might carry sensitive inform ation about attributes that users do not want to share. Various methods have bee n proposed to minimize the potential information leakage of sensitive attributes while maximizing the task accuracy. However, little is known about the theory b ehind these methods. In light of this gap, we develop a novel theoretical framew ork for attribute obfuscation. Under our framework, we propose a minimax optimiz ation formulation to protect the given attribute and analyze its inference guara ntees against worst-case adversaries. Meanwhile, there is a tension between mini mizing information leakage and maximizing task accuracy. To understand this, we prove an information-theoretic lower bound to precisely characterize the fundame ntal trade-off between accuracy and information leakage. We conduct experiments on two real-world datasets to corroborate the inference guarantees and validate the inherent trade-offs therein. Our results indicate that, among several altern atives, the adversarial learning approach achieves the best trade-off in terms o f attribute obfuscation and accuracy maximization.
**********

A Boolean Task Algebra for Reinforcement Learning
Geraud Nangue Tasse, Steven James, Benjamin Rosman
The ability to compose learned skills to solve new tasks is an important propert y for lifelong-learning agents. In this work we formalise the logical compositio n of tasks as a Boolean algebra. This allows us to formulate new tasks in terms of the negation, disjunction and conjunction of a set of base tasks. We then sho w that by learning goal-oriented value functions and restricting the transition dynamics of the tasks, an agent can solve these new tasks with no further learni ng. We prove that by composing these value functions in specific ways, we immedi ately recover the optimal policies for all tasks expressible under the Boolean a lgebra. We verify our approach in two domains---including a high-dimensional vid eo game environment requiring function approximation---where an agent first lear ns a set of base skills, and then composes them to solve a super-exponential num ber of new tasks.
**********

Learning with Differentiable Pertubed Optimizers
Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Ve rt, Francis Bach
Machine learning pipelines often rely on optimizers procedures to make discrete decisions (e.g., sorting, picking closest neighbors, or shortest paths). Althoug h these discrete decisions are easily computed in a forward manner, they break t he back-propagation of computational graphs. In order to expand the scope of lea rning problems that can be solved in an end-to-end fashion, we propose a systema tic method to transform optimizers into operations that are differentiable and n ever locally constant. Our approach relies on stochastically perturbed optimizer s, and can be used readily within existing solvers. Their derivatives can be eva luated efficiently, and smoothness tuned via the chosen noise amplitude. We also show how this framework can be connected to a family of losses developed in str uctured prediction, and give theoretical guarantees for their use in learning ta sks. We demonstrate experimentally the performance of our approach on various ta sks.
**********

Optimal Learning from Verified Training Data
Nicholas Bishop, Long Tran-Thanh, Enrico Gerding
Standard machine learning algorithms typically assume that data is sampled indep endently from the distribution of interest. In attempts to relax this assumption , fields such as adversarial learning typically assume that data is provided by

an adversary, whose sole objective is to fool a learning algorithm. However, in reality, it is often the case that data comes from self-interested agents, with less malicious goals and intentions which lie somewhere between the two settings described above. To tackle this problem, we present a Stackelberg competition model for least squares regression, in which data is provided by agents who wish to achieve specific predictions for their data. Although the resulting optimisation problem is nonconvex, we derive an algorithm which converges globally, outperforming current approaches which only guarantee convergence to local optima. We also provide empirical results on two real-world datasets, the medical personal costs dataset and the red wine dataset, showcasing the performance of our algorithm relative to algorithms which are optimal under adversarial assumptions, outperforming the state of the art.
**********

Online Linear Optimization with Many Hints
Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, Manish Purohit
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification
Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, Lenka Zdeborová
We analyze in a closed form the learning dynamics of stochastic gradient descent (SGD) for a single layer neural network classifying a high-dimensional Gaussian mixture where each cluster is assigned one of two labels. This problem provides a prototype of a non-convex loss landscape with interpolating regimes and a large generalization gap. We define a particular stochastic process for which SGD can be extended to a continuous-time limit that we call stochastic gradient flow. In the full-batch limit we recover the standard gradient flow. We apply dynamical mean-field theory from statistical physics to track the dynamics of the algorithm in the high-dimensional limit via a self-consistent stochastic process. We explore the performance of the algorithm as a function of control parameters shedding light on how it navigates the loss landscape.
**********

Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning
Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, Elias Bareinboim
One fundamental problem in the empirical sciences is of reconstructing the causal structure that underlies a phenomenon of interest through observation and experimentation. While there exists a plethora of methods capable of learning the equivalence class of causal structures that are compatible with observations, it is less well-understood how to systematically combine observations and experiments to reconstruct the underlying structure. In this paper, we investigate the task of structural learning in non-Markovian systems (i.e., when latent variables affect more than one observable) from a combination of observational and soft experimental data when the interventional targets are unknown. Using causal invariances found across the collection of observational and interventional distributions (not only conditional independences), we define a property called psi-Markov that connects these distributions to a pair consisting of (1) a causal graph D and (2) a set of interventional targets I. Building on this property, our main contributions are two-fold: First, we provide a graphical characterization that allows one to test whether two causal graphs with possibly different sets of interventional targets belong to the same psi-Markov equivalence class. Second, we develop an algorithm capable of harnessing the collection of data to learn the corresponding equivalence class. We then prove that this algorithm is sound and complete, in the sense that it is the most informative in the sample limit, i.e., it discovers as many tails and arrowheads as can be oriented within a psi-Markov equivalence class.
**********

Exploiting the Surrogate Gap in Online Multiclass Classification
Dirk van der Hoeven
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Pitfalls of Simplicity Bias in Neural Networks
Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, Praneeth Netrapalli
Several works have proposed Simplicity Bias (SB)---the tendency of standard training procedures such as Stochastic Gradient Descent (SGD) to find simple models---to justify why neural networks generalize well [Arpit et al. 2017, Nakkiran et al. 2019, Valle-Perez et al. 2019]. However, the precise notion of simplicity remains vague. Furthermore, previous settings [Soudry et al. 2018, Gunasekar et al. 2018] that use SB to theoretically justify why neural networks generalize well do not simultaneously capture the non-robustness of neural networks---a widely observed phenomenon in practice [Goodfellow et al. 2014, Jo and Bengio 2017]. We attempt to reconcile SB and the superior standard generalization of neural networks with the non-robustness observed in practice by introducing piecewise-linear and image-based datasets, which (a) incorporate a precise notion of simplicity, (b) comprise multiple predictive features with varying levels of simplicity, and (c) capture the non-robustness of neural networks trained on real data. Using theory and empirics on these datasets, we make four observations:
(i) SB of SGD and variants can be extreme: neural networks can exclusively rely on the simplest feature and remain invariant to all predictive complex features.
(ii) The extreme aspect of SB could explain why seemingly benign distribution shifts and small adversarial perturbations significantly degrade model performance.
(iii) Contrary to conventional wisdom, SB can also hurt generalization on the same data distribution, as SB persists even when the simplest feature has less predictive power than the more complex features.
(iv) Common approaches to improve generalization and robustness---ensembles and adversarial training---can fail in mitigating SB and its pitfalls.
Given the role of SB in training neural networks, we hope that the proposed datasets and methods serve as an effective testbed to evaluate novel algorithmic approaches aimed at avoiding the pitfalls of SB.
**********

Automatically Learning Compact Quality-aware Surrogates for Optimization Problems
Kai Wang, Bryan Wilder, Andrew Perrault, Milind Tambe
Solving optimization problems with unknown parameters often requires learning a predictive model to predict the values of the unknown parameters and then solving the problem using these values. Recent work has shown that including the optimization problem as a layer in the model training pipeline results in predictions of the unobserved parameters that lead to higher decision quality. Unfortunately, this process comes at a large computational cost because the optimization problem must be solved and differentiated through in each training iteration; furthermore, it may also sometimes fail to improve solution quality due to non-smoothness issues that arise when training through a complex optimization layer. To address these shortcomings, we learn a low-dimensional surrogate model of a large optimization problem by representing the feasible space in terms of meta-variables, each of which is a linear combination of the original variables. By training a low-dimensional surrogate model end-to-end, and jointly with the predictive model, we achieve: i) a large reduction in training and inference time; and ii) improved performance by focusing attention on the more important variables in the optimization and learning in a smoother space. Empirically, we demonstrate these improvements on a non-convex adversary modeling task, a submodular recommendation task and a convex portfolio optimization task.
**********

Empirical Likelihood for Contextual Bandits
Nikos Karampatziakis, John Langford, Paul Mineiro
We propose an estimator and confidence interval for computing the value of a policy from off-policy data in the contextual bandit setting. To this end we apply empirical likelihood techniques to formulate our estimator and confidence interval as simple convex optimization problems. Using the lower bound of our confidence interval, we then propose an off-policy policy optimization algorithm that searches for policies with large reward lower bound. We empirically find that both our estimator and confidence interval improve over previous proposals in finite sample regimes. Finally, the policy optimization algorithm we propose outperforms a strong baseline system for learning from off-policy data.
**********
Can Q-Learning with Graph Networks Learn a Generalizable Branching Heuristic for a SAT Solver?
Vitaly Kurin, Saad Godil, Shimon Whiteson, Bryan Catanzaro
We present Graph-Q-SAT, a branching heuristic for a Boolean SAT solver trained with value-based reinforcement learning (RL) using Graph Neural Networks for function approximation. Solvers using Graph-Q-SAT are complete SAT solvers that either provide a satisfying assignment or proof of unsatisfiability, which is required for many SAT applications. The branching heuristics commonly used in SAT solvers make poor decisions during their warm-up period, whereas Graph-Q-SAT is trained to examine the structure of the particular problem instance to make better decisions early in the search. Training Graph-Q-SAT is data efficient and does not require elaborate dataset preparation or feature engineering. We train Graph-Q-SAT using RL interfacing with MiniSat solver and show that Graph-Q-SAT can reduce the number of iterations required to solve SAT problems by 2-3X. Furthermore, it generalizes to unsatisfiable SAT instances, as well as to problems with 5X more variables than it was trained on. We show that for larger problems, reductions in the number of iterations lead to wall clock time reductions, the ultimate goal when designing heuristics.
We also show positive zero-shot transfer behavior when testing Graph-Q-SAT on a task family different from that used for training.
While more work is needed to apply Graph-Q-SAT to reduce wall clock time in modern SAT solving settings, it is a compelling proof-of-concept showing that RL equipped with Graph Neural Networks can learn a generalizable branching heuristic for SAT search.
**********
Non-reversible Gaussian processes for identifying latent dynamical structure in neural data
Virginia Rutten, Alberto Bernacchia, Maneesh Sahani, Guillaume Hennequin
A common goal in the analysis of neural data is to compress large population recordings into sets of interpretable, low-dimensional latent trajectories. This problem can be approached using Gaussian process (GP)-based methods which provide uncertainty quantification and principled model selection. However, standard GP priors do not distinguish between underlying dynamical processes and other forms of temporal autocorrelation. Here, we propose a new family of “dynamical” priors over trajectories, in the form of GP covariance functions that express a property shared by most dynamical systems: temporal non-reversibility. Non-reversibility is a universal signature of autonomous dynamical systems whose state trajectories follow consistent flow fields, such that any observed trajectory could not occur in reverse. Our new multi-output GP kernels can be used as drop-in replacements for standard kernels in multivariate regression, but also in latent variable models such as Gaussian process factor analysis (GPFA). We therefore introduce GPFADS (Gaussian Process Factor Analysis with Dynamical Structure), which models single-trial neural population activity using low-dimensional, non-reversible latent processes. Unlike previously proposed non-reversible multi-output kernels, ours admits a Kronecker factorization enabling fast and memory-efficient learning and inference. We apply GPFADS to synthetic data and show that it correctl

y recovers ground truth phase portraits. GPFADS also provides a probabilistic ge
neralization of jPCA, a method originally developed for identifying latent rotat
ional dynamics in neural data. When applied to monkey M1 neural recordings, GPFA
DS discovers latent trajectories with strong dynamical structure in the form of
rotations.
**********

Listening to Sounds of Silence for Speech Denoising
Ruilin Xu, Rundi Wu, Yuko Ishiwaka, Carl Vondrick, Changxi Zheng
We introduce a deep learning model for speech denoising, a long-standing challen
ge in audio analysis arising in numerous applications. Our approach is based on
a key observation about human speech: there is often a short pause between each
sentence or word. In a recorded speech signal, those pauses introduce a series o
f time periods during which only noise is present. We leverage these incidental
silent intervals to learn a model for automatic speech denoising given only mono
-channel audio. Detected silent intervals over time expose not just pure noise b
ut its time-varying features, allowing the model to learn noise dynamics and sup
press it from the speech signal. Experiments on multiple datasets confirm the pi
votal role of silent interval detection for speech denoising, and our method out
performs several state-of-the-art denoising methods, including those that accept
 only audio input (like ours) and those that denoise based on audiovisual input
(and hence require more information). We also show that our method enjoys excell
ent generalization properties, such as denoising spoken languages not seen durin
g training.
**********

BoxE: A Box Embedding Model for Knowledge Base Completion
Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, Tommaso Salvatori
Knowledge base completion (KBC) aims to automatically infer missing facts by exp
loiting information already present in a knowledge base (KB). A promising approa
ch for KBC is to embed knowledge into latent spaces and make predictions from le
arned embeddings.  However, existing embedding models are subject to at least on
e of the following limitations: (1) theoretical inexpressivity, (2) lack of supp
ort for prominent inference patterns (e.g., hierarchies), (3) lack of support fo
r KBC over higher-arity relations, and (4) lack of support for incorporating log
ical rules. Here, we propose a spatio-translational embedding model, called BoxE
, that simultaneously addresses all these limitations.  BoxE embeds entities as
points, and relations as a set of hyper-rectangles (or boxes), which spatially c
haracterize basic logical properties. This seemingly simple abstraction yields a
 fully expressive model offering a natural encoding for many desired logical pro
perties. BoxE can both capture and inject rules from rich classes of rule langua
ges, going well beyond individual inference patterns.  By design, BoxE naturally
 applies to higher-arity KBs. We conduct a detailed experimental analysis, and s
how that BoxE achieves state-of-the-art performance, both on benchmark knowledge
 graphs and on more general KBs, and we empirically show the power of integratin
g logical rules.
**********

Coherent Hierarchical Multi-Label Classification Networks
Eleonora Giunchiglia, Thomas Lukasiewicz
Hierarchical multi-label classification (HMC) is a challenging classification ta
sk extending standard multi-label classification problems by imposing a hierarch
y constraint on the classes.
In this paper, we propose C-HMCNN(h), a novel approach for HMC problems, which,
given a network h for the underlying multi-label classification problem, exploit
s the hierarchy information in order to produce  predictions coherent with the c
onstraint and improve performance.
We conduct an extensive experimental analysis
showing the superior performance of C-HMCNN(h) when compared to state-of-the-art
 models.
**********

Walsh-Hadamard Variational Inference for Bayesian Deep Learning
Simone Rossi, Sebastien Marmin, Maurizio Filippone

Over-parameterized models, such as DeepNets and ConvNets, form a class of models that are routinely adopted in a wide variety of applications, and for which Bayesian inference is desirable but extremely challenging. Variational inference offers the tools to tackle this challenge in a scalable way and with some degree of flexibility on the approximation, but for overparameterized models this is challenging due to the over-regularization property of the variational objective. Inspired by the literature on kernel methods, and in particular on structured approximations of distributions of random matrices, this paper proposes Walsh-Hadamard Variational Inference (WHVI), which uses Walsh-Hadamardbased factorization strategies to reduce the parameterization and accelerate computations, thus avoiding over-regularization issues with the variational objective. Extensive theoretical and empirical analyses demonstrate that WHVI yields considerable speedups and model reductions compared to other techniques to carry out approximate inference for over-parameterized models, and ultimately show how advances in kernel methods can be translated into advances in approximate Bayesian inference for Deep Learning.

**********

Federated Bayesian Optimization via Thompson Sampling

Zhongxiang Dai, Bryan Kian Hsiang Low, Patrick Jaillet

Bayesian optimization (BO) is a prominent approach to optimizing expensive-to-evaluate black-box functions. The massive computational capability of edge devices such as mobile phones, coupled with privacy concerns, has led to a surging interest in federated learning (FL) which focuses on collaborative training of deep neural networks (DNNs) via first-order optimization techniques. However, some common machine learning tasks such as hyperparameter tuning of DNNs lack access to gradients and thus require zeroth-order/black-box optimization. This hints at the possibility of extending BO to the FL setting (FBO) for agents to collaborate in these black-box optimization tasks. This paper presents federated Thompson sampling (FTS) which overcomes a number of key challenges of FBO and FL in a principled way: We (a) use random Fourier features to approximate the Gaussian process surrogate model used in BO, which naturally produces the parameters to be exchanged between agents, (b) design FTS based on Thompson sampling, which significantly reduces the number of parameters to be exchanged, and (c) provide a theoretical convergence guarantee that is robust against heterogeneous agents, which is a major challenge in FL and FBO. We empirically demonstrate the effectiveness of FTS in terms of communication efficiency, computational efficiency, and practical performance.

**********

MultiON: Benchmarking Semantic Map Memory using Multi-Object Navigation

Saim Wani, Shivansh Patel, Unnat Jain, Angel Chang, Manolis Savva

Navigation tasks in photorealistic 3D environments are challenging because they require perception and effective planning under partial observability. Recent work shows that map-like memory is useful for long-horizon navigation tasks. However, a focused investigation of the impact of maps on navigation tasks of varying complexity has not yet been performed.

**********

Neural Complexity Measures

Yoonho Lee, Juho Lee, Sung Ju Hwang, Eunho Yang, Seungjin Choi

While various complexity measures for deep neural networks exist, specifying an appropriate measure capable of predicting and explaining generalization in deep networks has proven challenging. We propose Neural Complexity (NC), a meta-learning framework for predicting generalization. Our model learns a scalar complexity measure through interactions with many heterogeneous tasks in a data-driven way. The trained NC model can be added to the standard training loss to regularize any task learner in a standard supervised learning scenario. We contrast NC's approach against existing manually-designed complexity measures and other meta-learning models, and we validate NC's performance on multiple regression and classification tasks.

**********

Optimal Iterative Sketching Methods with the Subsampled Randomized Hadamard Tran

sform
Jonathan Lacotte, Sifan Liu, Edgar Dobriban, Mert Pilanci
Random projections or sketching are widely used in many algorithmic and learning contexts. Here we study the performance of iterative Hessian sketch for least-squares problems. By leveraging and extending recent results from random matrix theory on the limiting spectrum of matrices randomly projected with the subsampled randomized Hadamard transform, and truncated Haar matrices, we can study and compare the resulting algorithms to a level of precision that has not been possible before. Our technical contributions include a novel formula for the second moment of the inverse of projected matrices. We also find simple closed-form expressions for asymptotically optimal step-sizes and convergence rates. These show that the convergence rate for Haar and randomized Hadamard matrices are identical, and asymptotically improve upon Gaussian random projections. These techniques may be applied to other algorithms that employ randomized dimension reduction.
**********
Provably adaptive reinforcement learning in metric spaces
Tongyi Cao, Akshay Krishnamurthy
We study reinforcement learning in continuous state and action spaces endowed with a metric. We provide a refined analysis of the algorithm of Sinclair, Banerjee, and Yu (2019) and show that its regret scales with the zooming dimension of the instance. This parameter, which originates in the bandit literature, captures the size of the subsets of near optimal actions and is always smaller than the covering dimension used in previous analyses. As such, our results are the first provably adaptive guarantees for reinforcement learning in metric spaces.
**********
ShapeFlow: Learnable Deformation Flows Among 3D Shapes
Chiyu Jiang, Jingwei Huang, Andrea Tagliasacchi, Leonidas J. Guibas
We present ShapeFlow, a flow-based model for learning a deformation space for entire classes of 3D shapes with large intra-class variations. ShapeFlow allows learning a multi-template deformation space that is agnostic to shape topology, yet preserves fine geometric details. Different from a generative space where a latent vector is directly decoded into a shape, a deformation space decodes a vector into a continuous flow that can advect a source shape towards a target. Such a space naturally allows the disentanglement of  geometric style (coming from the source) and structural pose (conforming to the target). We parametrize the deformation between geometries as a learned continuous flow field via a neural network and show that such deformations can be guaranteed to have desirable properties, such as bijectivity, freedom from self-intersections, or volume preservation. We illustrate the effectiveness of this learned deformation space for various downstream applications, including shape generation via deformation, geometric style transfer, unsupervised learning of a consistent parameterization for entire classes of shapes, and shape interpolation.
**********
Self-Supervised Learning by Cross-Modal Audio-Video Clustering
Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, Du Tran
Visual and audio modalities are highly correlated, yet they contain different information. Their strong correlation makes it possible to predict the semantics of one from the other with good accuracy. Their intrinsic differences make cross-modal prediction a potentially more rewarding pretext task for self-supervised learning of video and audio representations compared to within-modality learning. Based on this intuition, we propose Cross-Modal Deep Clustering (XDC), a novel self-supervised method that leverages unsupervised clustering in one modality (e.g., audio) as a supervisory signal for the other modality (e.g., video). This cross-modal supervision helps XDC utilize the semantic correlation and the differences between the two modalities. Our experiments show that XDC outperforms single-modality clustering and other multi-modal variants. XDC achieves state-of-the-art accuracy among self-supervised methods on multiple video and audio benchmarks. Most importantly, our video model pretrained on large-scale unlabeled data significantly outperforms the same model pretrained with full-supervision on Imag

eNet and Kinetics for action recognition on HMDB51 and UCF101. To the best of ou r knowledge, XDC is the first self-supervised learning method that outperforms l arge-scale fully-supervised pretraining for action recognition on the same archi tecture.
**********

Optimal Query Complexity of Secure Stochastic Convex Optimization
Wei Tang, Chien-Ju Ho, Yang Liu
We study the \emph{secure} stochastic convex optimization problem: a learner aim s to learn the optimal point of a convex function through sequentially querying a (stochastic) gradient oracle, in the meantime, there exists an adversary who a ims to free-ride and infer the learning outcome of the learner from observing th e learner's queries. The adversary observes only the points of the queries but n ot the feedback from the oracle.
The goal of the learner is to optimize the accuracy, i.e., obtaining an accurate estimate of the optimal point, while securing her privacy, i.e., making it diff icult for the adversary to infer the optimal point. We formally quantify this tr adeoff between learner’s accuracy and privacy and
characterize the lower and upper bounds on the learner's query complexity as a f unction of desired levels of accuracy and privacy. For the analysis of lower bou nds, we provide a general template based on information theoretical analysis and then tailor the template to several families of problems, including stochastic convex optimization and (noisy) binary search. We also present a generic secure learning protocol that achieves the matching upper bound up to logarithmic facto rs.
**********

DynaBERT: Dynamic BERT with Adaptive Width and Depth
Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, Qun Liu
The pre-trained language models like BERT, though powerful in many natural langu age processing tasks, are both computation and memory expensive. To alleviate th is problem, one approach is to compress them for specific tasks before deploymen t. However, recent works on BERT compression usually compress the large BERT mod el to a fixed smaller size, and can not fully satisfy the requirements of differ ent edge devices with various hardware performances. In this paper, we propose a novel dynamic BERT model (abbreviated as DynaBERT), which can flexibly adjust t he size and latency by selecting adaptive width and depth. The training process of DynaBERT includes first training a width-adaptive BERT and then allowing bot h adaptive width and depth, by distilling knowledge from the full-sized model to small sub-networks. Network rewiring is also used to keep the more important a ttention heads and neurons shared by more sub-networks. Comprehensive experimen ts under various efficiency constraints demonstrate that our proposed dynamic BE RT (or RoBERTa) at its largest size has comparable performance as BERT-base (or RoBERTa-base), while at smaller widths and depths consistently outperforms exist ing BERT compression methods.
Code is available at https://github.com/huawei-noah/Pretrained-Language-Model/tr ee/master/DynaBERT.
**********

Generalization Bound of Gradient Descent for Non-Convex Metric Learning
MINGZHI DONG, Xiaochen Yang, Rui Zhu, Yujiang Wang, Jing-Hao Xue
Metric learning aims to learn a distance measure that can benefit distance-based methods such as the nearest neighbour (NN) classifier. While considerable effor ts have been made to improve its empirical performance and analyze its generaliz ation ability by focusing on the data structure and model complexity, an unresol ved question is how choices of algorithmic parameters, such as the number of tra ining iterations, affect metric learning as it is typically formulated as an opt imization problem and nowadays more often as a non-convex problem. In this paper , we theoretically address this question and prove the agnostic Probably Approxi mately Correct (PAC) learnability for metric learning algorithms with non-convex objective functions optimized via gradient descent (GD); in particular, our the oretical guarantee takes the iteration number into account. We first show that t he generalization PAC bound is a sufficient condition for agnostic PAC learnabil

ity and this bound can be obtained by ensuring the uniform convergence on a dens ely concentrated subset of the parameter space. We then show that, for classifie rs optimized via GD, their generalizability can be guaranteed if the classifier and loss function are both Lipschitz smooth, and further improved by using fewer iterations. To illustrate and exploit the theoretical findings, we finally prop ose a novel metric learning method called Smooth Metric and representative Insta nce LEarning (SMILE), designed to satisfy the Lipschitz smoothness property and learned via GD with an early stopping mechanism for better discriminability and less computational cost of NN.
**********

Dynamic Submodular Maximization
Morteza Monemizadeh
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Inference for Batched Bandits
Kelly Zhang, Lucas Janson, Susan Murphy
As bandit algorithms are increasingly utilized in scientific studies and industr ial applications, there is an associated increasing need for reliable inference methods based on the resulting adaptively-collected data. In this work, we devel op methods for inference on data collected in batches using a bandit algorithm. We prove that the bandit arm selection probabilities cannot generally be assumed to concentrate. Non-concentration of the arm selection probabilities makes infe rence on adaptively-collected data challenging because classical statistical inf erence approaches, such as using asymptotic normality or the bootstrap, can have inflated Type-1 error and confidence intervals with below-nominal coverage prob abilities even asymptotically. In response we develop the Batched Ordinary Least Squares estimator (BOLS) that we prove is (1) asymptotically normal on data col lected from both multi-arm and contextual bandits and (2) robust to non-stationa rity in the baseline reward and thus leads to reliable Type-1 error control and accurate confidence intervals.
**********

Approximate Cross-Validation with Low-Rank Data in High Dimensions
Will Stephenson, Madeleine Udell, Tamara Broderick
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

GANSpace: Discovering Interpretable GAN Controls
Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, Sylvain Paris
This paper describes a simple technique to analyze Generative Adversarial Networ ks (GANs) and create interpretable controls for image synthesis, such as change of viewpoint, aging, lighting, and time of day. We identify important latent di rections based on Principal Component Analysis (PCA) applied either in latent sp ace or feature space. Then, we show that a large number of interpretable control s can be defined by layer-wise perturbation along the principal directions. Mor eover, we show that BigGAN can be controlled with layer-wise inputs in a StyleGA N-like manner. We show results on different GANs trained on various datasets, an d demonstrate good qualitative matches to edit directions found through earlier supervised approaches.
**********

Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bay esian Optimization
Samuel Daulton, Maximilian Balandat, Eytan Bakshy
In many real-world scenarios, decision makers seek to efficiently optimize multi ple competing objectives in a sample-efficient fashion. Multi-objective Bayesian optimization (BO) is a common approach, but many of the best-performing acquisi

tion functions do not have known analytic gradients and suffer from high computational overhead. We leverage recent advances in programming models and hardware acceleration for multi-objective BO using Expected Hypervolume Improvement (EHVI)---an algorithm notorious for its high computational complexity. We derive a novel formulation of q-Expected Hypervolume Improvement (qEHVI), an acquisition function that extends EHVI to the parallel, constrained evaluation setting. qEHVI is an exact computation of the joint EHVI of q new candidate points (up to Monte-Carlo (MC) integration error). Whereas previous EHVI formulations rely on gradient-free acquisition optimization or approximated gradients, we compute exact gradients of the MC estimator via auto-differentiation, thereby enabling efficient and effective optimization using first-order and quasi-second-order methods. Our empirical evaluation demonstrates that qEHVI is computationally tractable in many practical scenarios and outperforms state-of-the-art multi-objective BO algorithms at a fraction of their wall time.
**********

Neuron-level Structured Pruning using Polarization Regularizer
Tao Zhuang, Zhixuan Zhang, Yuheng Huang, Xiaoyi Zeng, Kai Shuang, Xiang 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.
**********

Limits on Testing Structural Changes in Ising Models
Aditya Gangrade, Bobak Nazer, Venkatesh Saligrama
We present novel information-theoretic limits on detecting sparse changes in Isingmodels, a problem that arises in many applications where network changes canoccur due to some external stimuli. We show that the sample complexity fordetecting sparse changes, in a minimax sense, is no better than learning the entiremodel even in settings with local sparsity. This is a surprising fact in light of priorwork rooted in sparse recovery methods, which suggest that sample complexityin this context scales only with the number of network changes. To shed light onwhen change detection is easier than structured learning, we consider testing ofedge deletion in forest-structured graphs, and high-temperature ferromagnets ascase studies. We show for these that testing of small changes is similarly hard, buttesting oflargechanges is well-separated from structure learning. These resultsimply that testing of graphical models may not be amenable to concepts such asrestricted strong convexity leveraged for sparsity pattern recovery, and algorithmdevelopment instead should be directed towards detection of large changes.
**********

Field-wise Learning for Multi-field Categorical Data
Zhibin Li, Jian Zhang, Yongshun Gong, Yazhou Yao, Qiang Wu
We propose a new method for learning with multi-field categorical data. Multi-field categorical data are usually collected over many heterogeneous groups. These groups can reflect in the categories under a field. The existing methods try to learn a universal model that fits all data, which is challenging and inevitably results in learning a complex model. In contrast, we propose a field-wise learning method leveraging the natural structure of data to learn simple yet efficient one-to-one field-focused models with appropriate constraints. In doing this, the models can be fitted to each category and thus can better capture the underlying differences in data. We present a model that utilizes linear models with variance and low-rank constraints, to help it generalize better and reduce the number of parameters. The model is also interpretable in a field-wise manner. As the dimensionality of multi-field categorical data can be very high, the models applied to such data are mostly over-parameterized. Our theoretical analysis can potentially explain the effect of over-parametrization on the generalization of our model. It also supports the variance constraints in the learning objective. The experiment results on two large-scale datasets show the superior performance of our model, the trend of the generalization error bound, and the interpretability of learning outcomes. Our code is available at https://github.com/lzb5600/Fie

ld-wise-Learning.
**********
Continual Learning in Low-rank Orthogonal Subspaces
Arslan Chaudhry, Naeemullah Khan, Puneet Dokania, Philip Torr
In continual learning (CL), a learner is faced with a sequence of tasks, arrivin g one after the other, and the goal is to remember all the tasks once the contin ual learning experience is finished. The prior art in CL uses episodic memory, p arameter regularization or extensible network structures to reduce interference among tasks, but in the end, all the approaches learn different tasks in a joint vector space. We believe this invariably leads to interference among different tasks. We propose to learn tasks in different (low-rank) vector subspaces that a re kept orthogonal to each other in order to minimize interference. Further, to keep the gradients of different tasks coming from these subspaces orthogonal to each other, we learn isometric mappings by posing network training as an optimiz ation problem over the Stiefel manifold. To the best of our understanding, we re port, for the first time, strong results over experience-replay baseline with an d without memory on standard classification benchmarks in continual learning.
**********
Unsupervised Learning of Visual Features by Contrasting Cluster Assignments
Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Arman d Joulin
Unsupervised image representations have significantly reduced the gap with super vised pretraining, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large num ber of explicit pairwise feature comparisons, which is computationally challengi ng. In this paper, we propose an online algorithm, SwAV, that takes advantage of contrastive methods without requiring to compute pairwise comparisons. Specific ally, our method simultaneously clusters the data while enforcing consistency be tween cluster assignments produced for different augmentations (or views) of the same image, instead of comparing features directly as in contrastive learning. Simply put, we use a swapped prediction mechanism where we predict the code of a view from the representation of another view. Our method can be trained with la rge and small batches and can scale to unlimited amounts of data. Compared to pr evious contrastive methods, our method is more memory efficient since it does no t require a large memory bank or a special momentum network. In addition, we als o propose a new data augmentation strategy, multi-crop, that uses a mix of views with different resolutions in place of two full-resolution views, without incre asing the memory or compute requirements. We validate our findings by achieving 75.3% top-1 accuracy on ImageNet with ResNet-50, as well as surpassing supervise d pretraining on all the considered transfer tasks.
**********
Sharpened Generalization Bounds based on Conditional Mutual Information and an A pplication to Noisy, Iterative Algorithms
Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, Gintare Karolina D ziugaite
The information-theoretic framework of Russo and Zou (2016) and Xu and Raginsky (2017) provides bounds on the generalization error of a learning algorithm in te rms of the mutual information between the algorithm's output and the training sa mple. In this work, we study the proposal, by Steinke and Zakynthinou (2020), to reason about the generalization error of a learning algorithm by introducing a super sample that contains the training sample as a random subset and computing mutual information conditional on the super sample. We first show that these new bounds based on the conditional mutual information are tighter than those based on the unconditional mutual information. We then introduce yet tighter bounds, building on the "individual sample" idea of Bu et al. (2019) and the "data depen dent" ideas of Negrea et al. (2019), using disintegrated mutual information. Fin ally, we apply these bounds to the study of Langevin dynamics algorithm, showing that conditioning on the super sample allows us to exploit information in the o ptimization trajectory to obtain tighter bounds based on hypothesis tests.
**********

Learning Deformable Tetrahedral Meshes for 3D Reconstruction
Jun Gao, Wenzheng Chen, Tommy Xiang, Alec Jacobson, Morgan McGuire, Sanja Fidler
3D shape representations that accommodate learning-based 3D reconstruction are a n open problem in machine learning and computer graphics. Previous work on neura l 3D reconstruction demonstrated benefits, but also limitations, of point cloud, voxel, surface mesh, and implicit function representations. We introduce \emph{ Deformable Tetrahedral Meshes} (DefTet) as a particular
parameterization that utilizes volumetric tetrahedral meshes for the reconstruct ion problem. Unlike existing volumetric approaches, DefTet optimizes for both vertex placement and occupancy, and is differentiable with respect to standard 3 D reconstruction loss functions. It is thus simultaneously high-precision, volum etric, and amenable to learning-based neural architectures. We show that it can represent arbitrary, complex topology, is both memory and computationally effici ent, and can produce high-fidelity reconstructions with a significantly smaller grid size than alternative volumetric approaches. The predicted surfaces are als o inherently defined as tetrahedral meshes, thus do not require post-processing. We demonstrate that DefTetmatches or exceeds both the quality of the previous be st approaches and the performance of the fastest ones. Our approach obtains high -quality tetrahedral meshes computed directly from noisy point clouds, and is th e first to showcase high-quality 3D results using only a single image as input.
**********

Information theoretic limits of learning a sparse rule
Clément Luneau, jean barbier, Nicolas Macris
We consider generalized linear models in regimes where the number of nonzero com ponents of the signal and accessible data points are sublinear with respect to t he size of the signal. We prove a variational formula for the asymptotic mutual information per sample when the system size grows to infinity. This result allow s us to derive an expression for the minimum mean-square error (MMSE) of the Bay esian estimator when the signal entries have a discrete distribution with finite support. We find that, for such signals and suitable vanishing scalings of the sparsity and sampling rate, the MMSE is nonincreasing piecewise constant. In spe cific instances the MMSE even displays an all-or-nothing phase transition, that is, the MMSE sharply jumps from its maximum value to zero at a critical sampling rate. The all-or-nothing phenomenon has previously been shown to occur in high- dimensional linear regression. Our analysis goes beyond the linear case and appl ies to learning the weights of a perceptron with general activation function in a teacher-student scenario. In particular, we discuss an all-or-nothing phenomen on for the generalization error with a sublinear set of training examples.
**********

Self-supervised learning through the eyes of a child
Emin Orhan, Vaibhav Gupta, Brenden M. Lake
Within months of birth, children develop meaningful expectations about the world around them. How much of this early knowledge can be explained through generic learning mechanisms applied to sensory data, and how much of it requires more substantive innate inductive biases? Addressing this fundamental question in its full generality is currently infeasible, but we can hope to make real progress in more narrowly defined domains, such as the development of high-level visual categories, thanks to improvements in data collecting technology and recent progress in deep learning. In this paper, our goal is precisely to achieve such progress by utilizing modern self-supervised deep learning methods and a recent longitudinal, egocentric video dataset recorded from the perspective of three young children (Sullivan et al., 2020). Our results demonstrate the emergence of powerful, high-level visual representations from developmentally realistic natural videos using generic self-supervised learning objectives.
**********

Unsupervised Semantic Aggregation and Deformable Template Matching for Semi-Supe rvised Learning
Tao Han, Junyu Gao, Yuan Yuan, Qi Wang
Unlabeled data learning has attracted considerable attention recently. However,

it is still elusive to extract the expected high-level semantic feature with mere unsupervised learning. In the meantime, semi-supervised learning (SSL) demonstrates a promising future in leveraging few samples. In this paper, we combine both to propose an Unsupervised Semantic Aggregation and Deformable Template Matching (USADTM) framework for SSL, which strives to improve the classification performance with few labeled data and then reduce the cost in data annotating. Specifically, unsupervised semantic aggregation based on Triplet Mutual Information (T-MI) loss is explored to generate semantic labels for unlabeled data. Then the semantic labels are aligned to the actual class by the supervision of labeled data. Furthermore, a feature pool that stores the labeled samples is dynamically updated to assign proxy labels for unlabeled data, which are used as targets for cross-entropy minimization. Extensive experiments and analysis across four standard semi-supervised learning benchmarks validate that USADTM achieves top performance (e.g., 90.46% accuracy on CIFAR-10 with 40 labels and 95.20% accuracy with 250 labels). The code is released at https://github.com/taohan10200/USADTM.
**********
A game-theoretic analysis of networked system control for common-pool resource management using multi-agent reinforcement learning
Arnu Pretorius, Scott Cameron, Elan van Biljon, Thomas Makkink, Shahil Mawjee, Jeremy du Plessis, Jonathan Shock, Alexandre Laterre, Karim Beguir
Multi-agent reinforcement learning has recently shown great promise as an approach to networked system control. Arguably, one of the most difficult and important tasks for which large scale networked system control is applicable is common-pool resource management. Crucial common-pool resources include arable land, fresh water, wetlands, wildlife, fish stock, forests and the atmosphere, of which proper management is related to some of society's greatest challenges such as food security, inequality and climate change. Here we take inspiration from a recent research program investigating the game-theoretic incentives of humans in social dilemma situations such as the well-known \textit{tragedy of the commons}. However, instead of focusing on biologically evolved human-like agents, our concern is rather to better understand the learning and operating behaviour of engineered networked systems comprising general-purpose reinforcement learning agents, subject only to nonbiological constraints such as memory, computation and communication bandwidth. Harnessing tools from empirical game-theoretic analysis, we analyse the differences in resulting solution concepts that stem from employing different information structures in the design of networked multi-agent systems. These information structures pertain to the type of information shared between agents as well as the employed communication protocol and network topology. Our analysis contributes new insights into the consequences associated with certain design choices and provides an additional dimension of comparison between systems beyond efficiency, robustness, scalability and mean control performance.
**********
What shapes feature representations? Exploring datasets, architectures, and training
Katherine Hermann, Andrew Lampinen
In naturalistic learning problems, a model's input contains a wide range of features, some useful for the task at hand, and others not. Of the useful features, which ones does the model use? Of the task-irrelevant features, which ones does the model represent? Answers to these questions are important for understanding the basis of models' decisions, as well as for building models that learn versatile, adaptable representations useful beyond the original training task. We study these questions using synthetic datasets in which the task-relevance of input features can be controlled directly. We find that when two features redundantly predict the labels, the model preferentially represents one, and its preference reflects what was most linearly decodable from the untrained model. Over training, task-relevant features are enhanced, and task-irrelevant features are partially suppressed. Interestingly, in some cases, an easier, weakly predictive feature can suppress a more strongly predictive, but more difficult one. Additionally, models trained to recognize both easy and hard features learn representations most similar to models that use only the easy feature. Further, easy features lea

d to more consistent representations across model runs than do hard features. Fi
nally, models have greater representational similarity to an untrained model tha
n to models trained on a different task. Our results highlight the complex proce
sses that determine which features a model represents.
**********

Optimal Best-arm Identification in Linear Bandits
Yassir Jedra, Alexandre Proutiere
We study the problem of best-arm identification with fixed confidence in stochas
tic linear bandits. The objective is to identify the best arm with a given level
 of certainty while minimizing the sampling budget. We devise a simple algorithm
 whose sampling complexity matches known instance-specific lower bounds, asympto
tically almost surely and in expectation. The algorithm relies on an arm samplin
g rule that tracks an optimal proportion of arm draws, and that remarkably can b
e updated as rarely as we wish, without compromising its theoretical guarantees.
 Moreover, unlike existing best-arm identification strategies, our algorithm use
s a stopping rule that does not depend on the number of arms. Experimental resul
ts suggest that our algorithm significantly outperforms existing algorithms. The
 paper further provides a first analysis of the best-arm identification problem
in linear bandits with a continuous set of arms.
**********

Data Diversification: A Simple Strategy For Neural Machine Translation
Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, Ai Ti Aw
We introduce Data Diversification: a simple but effective strategy to boost neur
al machine translation (NMT) performance. It diversifies the training data by us
ing the predictions of multiple forward and backward models and then merging the
m with the original dataset on which the final NMT model is trained. Our method
is applicable to all NMT models. It does not require extra monolingual data like
 back-translation, nor does it add more computations and parameters like ensembl
es of models. Our method achieves state-of-the-art BLEU scores of 30.7 and 43.7
in the WMT'14 English-German and English-French translation tasks, respectively.
 It also substantially improves on 8 other translation tasks: 4 IWSLT tasks (Eng
lish-German and English-French) and 4 low-resource translation tasks (English-Ne
pali and English-Sinhala). We demonstrate that our method is more effective than
 knowledge distillation and dual learning, it exhibits strong correlation with e
nsembles of models, and it trades perplexity off for better BLEU score.
**********

Interstellar: Searching Recurrent Architecture for ■Knowledge Graph Embedding
Yongqi Zhang, Quanming Yao, Lei  Chen
Knowledge graph (KG) embedding is well-known in learning representations of KGs.
 Many models have been proposed to learn the interactions between entities and r
elations of the triplets. However, long-term information among multiple triplets
 is also important to KG. In this work, based on the relational paths, which are
 composed of a sequence of triplets, we define the Interstellar as a recurrent n
eural architecture search problem for the short-term and long-term information a
long the paths. First, we analyze the difficulty of using a unified model to wor
k as the Interstellar. Then, we propose to search for recurrent architecture as
the Interstellar for different KG tasks. A case study on synthetic data illustra
tes the importance of the defined search problem. Experiments on real datasets d
emonstrate the effectiveness of the searched models and the efficiency of the pr
oposed hybrid-search algorithm.
**********

CoSE: Compositional Stroke Embeddings
Emre Aksan, Thomas Deselaers, Andrea Tagliasacchi, Otmar Hilliges
We present a generative model for stroke-based drawing tasks which is able to mo
del complex free-form structures. While previous approaches rely on sequence-bas
ed models for drawings of basic objects or handwritten text, we propose a model
that treats drawings as a collection of strokes that can be composed into comple
x structures such as diagrams (e.g., flow-charts). At the core of the approach l
ies a novel auto-encoder that projects variable-length strokes into a latent spa
ce of fixed dimension. This representation space allows a relational model, oper

ating in latent space, to better capture the relationship between strokes and to predict subsequent strokes. We demonstrate qualitatively and quantitatively that our proposed approach is able to model the appearance of individual strokes, as well as the compositional structure of larger diagram drawings. Our approach is suitable for interactive use cases such as auto-completing diagrams. We make code and models publicly available at https://eth-ait.github.io/cose.
**********
Learning Multi-Agent Coordination for Enhancing Target Coverage in Directional Sensor Networks
Jing Xu, Fangwei Zhong, Yizhou Wang
Maximum target coverage by adjusting the orientation of distributed sensors is an important problem in directional sensor networks (DSNs). This problem is challenging as the targets usually move randomly but the coverage range of sensors is limited in angle and distance. Thus, it is required to coordinate sensors to get ideal target coverage with low power consumption, e.g. no missing targets or reducing redundant coverage. To realize this, we propose a Hierarchical Target-oriented Multi-Agent Coordination (HiT-MAC), which decomposes the target coverage problem into two-level tasks: targets assignment by a coordinator and tracking assigned targets by executors. Specifically, the coordinator periodically monitors the environment globally and allocates targets to each executor. In turn, the executor only needs to track its assigned targets. To effectively learn the HiT-MAC by reinforcement learning, we further introduce a bunch of practical methods, including a self-attention module, marginal contribution approximation for the coordinator, goal-conditional observation filter for the executor, etc. Empirical results demonstrate the advantage of HiT-MAC in coverage rate, learning efficiency, and scalability, comparing to baselines. We also conduct an ablative analysis on the effectiveness of the introduced components in the framework.
**********
Biological credit assignment through dynamic inversion of feedforward networks
Bill Podlaski, Christian K. Machens
Learning depends on changes in synaptic connections deep inside the brain. In multilayer networks, these changes are triggered by error signals fed back from the output, generally through a stepwise inversion of the feedforward processing steps. The gold standard for this process --- backpropagation --- works well in artificial neural networks, but is biologically implausible. Several recent proposals have emerged to address this problem, but many of these biologically-plausible schemes are based on learning an independent set of feedback connections. This complicates the assignment of errors to each synapse by making it dependent upon a second learning problem, and by fitting inversions rather than guaranteeing them. Here, we show that feedforward network transformations can be effectively inverted through dynamics. We derive this dynamic inversion from the perspective of feedback control, where the forward transformation is reused and dynamically interacts with fixed or random feedback to propagate error signals during the backward pass. Importantly, this scheme does not rely upon a second learning problem for feedback because accurate inversion is guaranteed through the network dynamics. We map these dynamics onto generic feedforward networks, and show that the resulting algorithm performs well on several supervised and unsupervised datasets. Finally, we discuss potential links between dynamic inversion and second-order optimization. Overall, our work introduces an alternative perspective on credit assignment in the brain, and proposes a special role for temporal dynamics and feedback control during learning.
**********
Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching
Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, Dejing Dou
Discriminatively localizing sounding objects in cocktail-party, i.e., mixed sound scenes, is commonplace for humans, but still challenging for machines. In this paper, we propose a two-stage learning framework to perform self-supervised class-aware sounding object localization. First, we propose to learn robust object

representations by aggregating the candidate sound localization results in the s
ingle source scenes. Then, class-aware object localization maps are generated in
 the cocktail-party scenarios by referring the pre-learned object knowledge, and
 the sounding objects are accordingly selected by matching audio and visual obje
ct category distributions, where the audiovisual consistency is viewed as the se
lf-supervised signal. Experimental results in both realistic and synthesized coc
ktail-party videos demonstrate that our model is superior in filtering out silen
t objects and pointing out the location of sounding objects of different classes
. Code is available at https://github.com/DTaoo/Discriminative-Sounding-Objects-
Localization.
**********

Learning Multi-Agent Communication through Structured Attentive Reasoning
Murtaza Rangwala, Ryan Williams
Learning communication via deep reinforcement learning has recently been shown t
o be an effective way to solve cooperative multi-agent tasks. However, learning
 which communicated information is beneficial for each agent's decision-making pr
ocess remains a challenging task. In order to address this problem, we explore r
elational reinforcement learning which leverages attention-based networks to lea
rn efficient and interpretable relations between entities. On the foundation of
relations, we introduce a novel communication architecture that exploits a memor
y-based attention network that selectively reasons about the value of informatio
n received from other agents while considering its past experiences.  Specifical
ly, the model communicates by first computing the relevance of messages received
 from other agents and then extracts task-relevant information from memories giv
en the newly received information. We empirically demonstrate the strength of ou
r model in cooperative and competitive multi-agent tasks, where inter-agent comm
unication and reasoning over prior information substantially improves performanc
e compared to baselines. We further show in the accompanying videos and experime
ntal results that the agents learn a sophisticated and diverse set of cooperativ
e behaviors to solve challenging tasks, both for discrete and continuous action
spaces using on-policy and off-policy gradient methods. By developing an explici
t architecture that is targeted towards communication, our work aims to open new
 directions to overcome important challenges in multi-agent cooperation through
learned communication.
**********

Private Identity Testing for High-Dimensional Distributions
Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, Lydia Zakynt
hinou
In this work we present novel differentially private identity (goodness-of-fit)
testers for natural and widely studied classes of multivariate product distribut
ions: Gaussians in R^d with known covariance and product distributions over {\pm
 1}^d. Our testers have improved sample complexity compared to those derived fro
m previous techniques, and are the first testers whose sample complexity matches
 the order-optimal minimax sample complexity of O(d^1/2/alpha^2) in many paramet
er regimes. We construct two types of testers, exhibiting tradeoffs between samp
le complexity and computational complexity. Finally, we provide a two-way reduct
ion between testing a subclass of multivariate product distributions and testing
 univariate distributions, and thereby obtain upper and lower bounds for testing
 this subclass of product distributions.
**********

On the Optimal Weighted $\ell_2$ Regularization in Overparameterized Linear Regr
ession
Denny Wu, Ji 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

An Efficient Asynchronous Method for Integrating Evolutionary and Gradient-based
 Policy Search

Kyunghyun Lee, Byeong-Uk Lee, Ukcheol Shin, In So Kweon
Deep reinforcement learning (DRL) algorithms and evolution strategies (ES) have been applied to various tasks, showing excellent performances.
These have the opposite properties, with DRL having good sample efficiency and poor stability, while ES being vice versa.
Recently, there have been attempts to combine these algorithms, but these methods fully rely on synchronous update scheme, making it not ideal to maximize the benefits of the parallelism in ES.
To solve this challenge, asynchronous update scheme was introduced, which is capable of good time-efficiency and diverse policy exploration.
In this paper, we introduce an Asynchronous Evolution Strategy-Reinforcement Learning (AES-RL) that maximizes the parallel efficiency of ES and integrates it with policy gradient methods.
Specifically, we propose 1) a novel framework to merge ES and DRL asynchronously and 2) various asynchronous update methods that can take all advantages of asynchronism, ES, and DRL, which are exploration and time efficiency, stability, and sample efficiency, respectively.
The proposed framework and update methods are evaluated in continuous control benchmark work, showing superior performance as well as time efficiency compared to the previous methods.
**********

MetaSDF: Meta-Learning Signed Distance Functions
Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, Gordon Wetzstein
Neural implicit shape representations are an emerging paradigm that offers many potential benefits over conventional discrete representations, including memory efficiency at a high spatial resolution. Generalizing across shapes with such neural implicit representations amounts to learning priors over the respective function space and enables geometry reconstruction from partial or noisy observations. Existing generalization methods rely on conditioning a neural network on a low-dimensional latent code that is either regressed by an encoder or jointly optimized in the auto-decoder framework. Here, we formalize learning of a shape space as a meta-learning problem and leverage gradient-based meta-learning algorithms to solve this task. We demonstrate that this approach performs on par with auto-decoder based approaches while being an order of magnitude faster at test-time inference. We further demonstrate that the proposed gradient-based method outperforms encoder-decoder based methods that leverage pooling-based set encoders.
**********

Simple and Scalable Sparse k-means Clustering via Feature Ranking
Zhiyue Zhang, Kenneth Lange, Jason Xu
Clustering, a fundamental activity in unsupervised learning, is notoriously difficult when the feature space is high-dimensional. Fortunately, in many realistic scenarios, only a handful of features are relevant in distinguishing clusters. This has motivated the development of sparse clustering techniques that typically rely on k-means within outer algorithms of high computational complexity. Current techniques also require careful tuning of shrinkage parameters, further limiting their scalability. In this paper, we propose a novel framework for sparse k-means clustering that is intuitive, simple to implement, and competitive with state-of-the-art algorithms. We show that our algorithm enjoys consistency and convergence guarantees. Our core method readily generalizes to several task-specific algorithms such as clustering on subsets of attributes and in partially observed data settings. We showcase these contributions thoroughly via simulated experiments and real data benchmarks, including a case study on protein expression in trisomic mice.
**********

Model-based Adversarial Meta-Reinforcement Learning
Zichuan Lin, Garrett Thomas, Guangwen Yang, Tengyu Ma
Meta-reinforcement learning (meta-RL) aims to learn from multiple training tasks the ability to adapt efficiently to unseen test tasks. Despite the success, existing meta-RL algorithms are known to be sensitive to the task distribution shift. When the test task distribution is different from the training task distribut

ion, the performance may degrade significantly. To address this issue, this pape r proposes \textit{Model-based Adversarial Meta-Reinforcement Learning} (AdMRL), where we aim to minimize the worst-case sub-optimality gap --- the difference b etween the optimal return and the return that the algorithm achieves after adapt ation --- across all tasks in a family of tasks, with a model-based approach. W e propose a minimax objective and optimize it by alternating between learning th e dynamics model on a fixed task and finding the \textit{adversarial} task for t he current model --- the task for which the policy induced by the model is maxim ally suboptimal. Assuming the family of tasks is parameterized, we derive a for mula for the gradient of the suboptimality with respect to the task parameters v ia the implicit function theorem, and show how the gradient estimator can be eff iciently implemented by the conjugate gradient method and a novel use of the REI NFORCE estimator. We evaluate our approach on several continuous control benchma rks and demonstrate its efficacy in the worst-case performance over all tasks, t he generalization power to out-of-distribution tasks, and in training and test t ime sample efficiency, over existing state-of-the-art meta-RL algorithms.
**********

Graph Policy Network for Transferable Active Learning on Graphs
Shengding Hu, Zheng Xiong, Meng Qu, Xingdi Yuan, Marc-Alexandre Côté, Zhiyuan Li u, Jian Tang
Graph neural networks (GNNs) have been attracting increasing popularity due to t heir simplicity and effectiveness in a variety of fields. However, a large numbe r of labeled data is generally required to train these networks, which could be very expensive to obtain in some domains. In this paper, we study active learnin g for GNNs, i.e., how to efficiently label the nodes on a graph to reduce the an notation cost of training GNNs. We formulate the problem as a sequential decisio n process on graphs and train a GNN-based policy network with reinforcement lear ning to learn the optimal query strategy. By jointly training on several source graphs with full labels, we learn a transferable active learning policy which ca n directly generalize to unlabeled target graphs. Experimental results on multip le datasets from different domains prove the effectiveness of the learned policy in promoting active learning performance in both settings of transferring betwe en graphs in the same domain and across different domains.
**********

Towards a Better Global Loss Landscape of GANs
Ruoyu Sun, Tiantian Fang, Alexander Schwing
Understanding of GAN training is still very limited. One major challenge is its non-convex-non-concave min-max objective, which may lead to sub-optimal local m inima. In this work, we perform a global landscape analysis of the empirical los s of GANs. We prove that a class of separable-GAN, including the original JS-GAN , has exponentially many bad basins which are perceived as mode-collapse. We als o study the relativistic pairing GAN (RpGAN) loss which couples the generated sa mples and the true samples. We prove that RpGAN has no bad basins. Experiments o n synthetic data show that the predicted bad basin can indeed appear in trainin g. We also perform experiments to support our theory that RpGAN has a better la ndscape than separable-GAN. For instance, we empirically show that RpGAN perform s better than separable-GAN with relatively narrow neural nets. The code is avai lable at \url{https://github.com/AilsaF/RS-GAN}.
**********

Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-A gent Reinforcement Learning
Tabish Rashid, Gregory Farquhar, Bei Peng, Shimon Whiteson
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

BanditPAM: Almost Linear Time k-Medoids Clustering via Multi-Armed Bandits
Mo Tiwari, Martin J. Zhang, James Mayclin, Sebastian Thrun, Chris Piech, Ilan Sh omorony

Clustering is a ubiquitous task in data science. Compared to the commonly used k-means clustering, k-medoids clustering requires the cluster centers to be actual data points and supports arbitrary distance metrics, which permits greater interpretability and the clustering of structured objects. Current state-of-the-art k-medoids clustering algorithms, such as Partitioning Around Medoids (PAM), are iterative and are quadratic in the dataset size n for each iteration, being prohibitively expensive for large datasets. We propose BanditPAM, a randomized algorithm inspired by techniques from multi-armed bandits, that reduces the complexity of each PAM iteration from O(n^2) to O(nlogn) and returns the same results with high probability, under assumptions on the data that often hold in practice. As such, BanditPAM matches state-of-the-art clustering loss while reaching solutions much faster. We empirically validate our results on several large real-world datasets, including a coding exercise submissions dataset from Code.org, the 10x Genomics 68k PBMC single-cell RNA sequencing dataset, and the MNIST handwritten digits dataset. In these experiments, we observe that BanditPAM returns the same results as state-of-the-art PAM-like algorithms up to 4x faster while performing up to 200x fewer distance computations. The improvements demonstrated by BanditPAM enable k-medoids clustering on a wide range of applications, including identifying cell types in large-scale single-cell data and providing scalable feedback for students learning computer science online. We also release highly optimized Python and C++ implementations of our algorithm.

**********

UDH: Universal Deep Hiding for Steganography, Watermarking, and Light Field Messaging

Chaoning Zhang, Philipp Benz, Adil Karjauv, Geng Sun, In So Kweon

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

**********

Evidential Sparsification of Multimodal Latent Spaces in Conditional Variational Autoencoders

Masha Itkina, Boris Ivanovic, Ransalu Senanayake, Mykel J. Kochenderfer, Marco Pavone

Discrete latent spaces in variational autoencoders have been shown to effectively capture the data distribution for many real-world problems such as natural language understanding, human intent prediction, and visual scene representation. However, discrete latent spaces need to be sufficiently large to capture the complexities of real-world data, rendering downstream tasks computationally challenging. For instance, performing motion planning in a high-dimensional latent representation of the environment could be intractable. We consider the problem of sparsifying the discrete latent space of a trained conditional variational autoencoder, while preserving its learned multimodality. As a post hoc latent space reduction technique, we use evidential theory to identify the latent classes that receive direct evidence from a particular input condition and filter out those that do not. Experiments on diverse tasks, such as image generation and human behavior prediction, demonstrate the effectiveness of our proposed technique at reducing the discrete latent sample space size of a model while maintaining its learned multimodality.

**********

An Unbiased Risk Estimator for Learning with Augmented Classes

Yu-Jie Zhang, Peng Zhao, Lanjihong Ma, Zhi-Hua Zhou

This paper studies the problem of learning with augmented classes (LAC), where augmented classes unobserved in the training data might emerge in the testing phase. Previous studies generally attempt to discover augmented classes by exploiting geometric properties, achieving inspiring empirical performance yet lacking theoretical understandings particularly on the generalization ability. In this paper we show that, by using unlabeled training data to approximate the potential distribution of augmented classes, an unbiased risk estimator of the testing distribution can be established for the LAC problem under mild assumptions, which p

aves a way to develop a sound approach with theoretical guarantees. Moreover, th e proposed approach can adapt to complex changing environments where augmented c lasses may appear and the prior of known classes may change simultaneously. Exte nsive experiments confirm the effectiveness of our proposed approach.
**********

AutoBSS: An Efficient Algorithm for Block Stacking Style Search
Yikang Zhang, Jian Zhang, Zhao Zhong
Neural network architecture design mostly focuses on the new convolutional opera tor or special topological structure of network block, little attention is drawn to the configuration of stacking each block, called Block Stacking Style (BSS). Recent studies show that BSS may also have an unneglectable impact on networks, thus we design an efficient algorithm to search it automatically. The proposed method, AutoBSS, is a novel AutoML algorithm based on Bayesian optimization by i teratively refining and clustering Block Stacking Style Code (BSSC), which can f ind optimal BSS in a few trials without biased evaluation. On ImageNet classific ation task, ResNet50/MobileNetV2/EfficientNet-B0 with our searched BSS achieve 7 9.29%/74.5%/77.79%, which outperform the original baselines by a large margin. M ore importantly, experimental results on model compression, object detection and instance segmentation show the strong generalizability of the proposed AutoBSS, and further verify the unneglectable impact of BSS on neural networks.
**********

Pushing the Limits of Narrow Precision Inferencing at Cloud Scale with Microsoft Floating Point
Bita Darvish Rouhani, Daniel Lo, Ritchie Zhao, Ming Liu, Jeremy Fowers, Kalin Ov tcharov , Anna Vinogradsky, Sarah Massengill , Lita Yang, Ray Bittner, Alessandr o Forin, Haishan Zhu, Taesik Na, Prerak Patel, Shuai Che, Lok Chand Koppaka , XI A SONG, Subhojit  Som, Kaustav  Das, Saurabh T, Steve Reinhardt , Sitaram Lanka, Eric Chung, Doug Burger
In this paper, we explore the limits of Microsoft Floating Point (MSFP), a new c lass of datatypes developed for production cloud-scale inferencing on custom har dware. Through the co-evolution of hardware design and algorithms, MSFP achieves accuracy comparable to or better than industry standards Bfloat16 and INT8 at 3 x and 4x lower cost, respectively. MSFP incurs negligible impact to accuracy (<1 %), requires no changes to the model topology, and is integrated with a mature c loud production pipeline. MSFP supports various classes of deep learning models including CNNs, RNNs, and Transformers without modification. Finally, we charact erize the accuracy and implementation of MSFP and demonstrate its efficacy on a number of production scenarios, including models that power major online scenari os such as web search, question-answering, and image classification.
**********

Stochastic Optimization with Laggard Data Pipelines
Naman Agarwal, Rohan Anil, Tomer Koren, Kunal Talwar, Cyril Zhang
State-of-the-art optimization is steadily shifting towards massively parallel pi pelines with extremely large batch sizes. As a consequence, CPU-bound preprocess ing and disk/memory/network operations have emerged as new performance bottlenec ks, as opposed to hardware-accelerated gradient computations. In this regime, a recently proposed approach is data echoing (Choi et al., 2019), which takes repe ated gradient steps on the same batch while waiting for fresh data to arrive fro m upstream. We provide the first convergence analyses of "data-echoed" extension s of common optimization methods, showing that they exhibit provable improvement s over their synchronous counterparts. Specifically, we show that in convex opti mization with stochastic minibatches, data echoing affords speedups on the curva ture-dominated part of the convergence rate, while maintaining the optimal stati stical rate.
**********

Self-supervised Auxiliary Learning with Meta-paths for Heterogeneous Graphs
Dasol Hwang, Jinyoung Park, Sunyoung Kwon, KyungMin Kim, Jung-Woo Ha, Hyunwoo J. Kim
Graph neural networks have shown superior performance in a wide range of applica tions providing a powerful representation of graph-structured data.  Recent work

s show that the representation can be further improved by auxiliary tasks. Howev
er, the auxiliary tasks for heterogeneous graphs, which contain rich semantic in
formation with various types of nodes and edges, have less explored in the liter
ature. In this paper, to learn graph neural networks on heterogeneous graphs we
propose a novel self-supervised auxiliary learning method using meta paths, whic
h are composite relations of multiple edge types. Our proposed method is learnin
g to learn a primary task by predicting meta-paths as auxiliary tasks. This can
be viewed as a type of meta-learning. The proposed method can identify an effect
ive combination of auxiliary tasks and automatically balance them to improve the
 primary task. Our methods can be applied to any graph neural networks in a plug
-in manner without manual labeling or additional data. The experiments demonstra
te that the proposed method consistently improves the performance of link predic
tion and node classification on heterogeneous graphs.
**********

GPS-Net: Graph-based Photometric Stereo Network
Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, Boxin Shi
Learning-based photometric stereo methods predict the surface normal either in a
 per-pixel or an all-pixel manner. Per-pixel methods explore the inter-image int
ensity variation of each pixel but ignore features from the intra-image spatial
domain. All-pixel methods explore the intra-image intensity variation of each in
put image but pay less attention to the inter-image lighting variation. In this
paper, we present a Graph-based Photometric Stereo Network, which unifies per-pi
xel and all-pixel processings to explore both inter-image and intra-image inform
ation. For per-pixel operation, we propose the Unstructured Feature Extraction L
ayer to connect an arbitrary number of input image-light pairs into graph struct
ures, and introduce Structure-aware Graph Convolution filters to balance the inp
ut data by appropriately weighting shadows and specular highlights. For all-pixe
l operation, we propose the Normal Regression Network to make efficient use of t
he intra-image spatial information for predicting a surface normal map with rich
 details. Experimental results on the real-world benchmark show that our method
achieves excellent performance under both sparse and dense lighting distribution
s.
**********

Consistent Structural Relation Learning for Zero-Shot Segmentation
Peike Li, Yunchao Wei, Yi Yang
Zero-shot semantic segmentation aims to recognize the semantics of pixels from u
nseen categories with zero training samples. Previous practice [1] proposed to t
rain the classifiers for unseen categories using the visual features generated f
rom semantic word embeddings. However, the generator is merely learned on the se
en categories while no constraint is applied to the unseen categories, leading t
o poor generalization ability.  In this work, we propose a Consistent Structural
 Relation Learning (CSRL) approach to constrain the generating of unseen visual
features by exploiting the structural relations between seen and unseen categori
es. We observe that different categories are usually with similar relations in e
ither semantic word embedding space or visual feature space. This observation mo
tivates us to harness the similarity of category-level relations on the semantic
 word embedding space to learn a better visual feature generator. Concretely, by
 exploring the pair-wise and list-wise structures, we impose the relations of ge
nerated visual features to be consistent with their counterparts in the semantic
 word embedding space. In this way, the relations between seen and unseen catego
ries will be transferred to implicitly constrain the generator to produce relati
on-consistent unseen visual features. We conduct extensive experiments on Pascal
-VOC and Pascal-Context benchmarks. The proposed CSRL significantly outperforms
existing state-of-the-art methods by a large margin, resulting in ~7-12% on Pasc
al-VOC and ~2-5% on Pascal-Context.
**********

Model Selection in Contextual Stochastic Bandit Problems
Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lat
timore, Csaba Szepesvari
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Truncated Linear Regression in High Dimensions
Constantinos Daskalakis, Dhruv Rohatgi, Emmanouil Zampetakis
As in standard linear regression, in truncated linear regression, we are given a ccess to observations (Ai, yi)i whose dependent variable equals yi= Ai^{\rm T} \ cdot x^* + \etai, where x^* is some fixed unknown vector of interest and \etai i s independent noise; except we are only given an observation if its dependent va riable yi lies in some "truncation set" S \subset \mathbb{R}. The goal is to rec over x^* under some favorable conditions on the Ai's and the noise distribution. We prove that there exists a computationally and statistically efficient method for recovering k-sparse n-dimensional vectors x^* from m truncated samples, whi ch attains an optimal \ell2 reconstruction error of O(\sqrt{(k \log n)/m}). As a corollary, our guarantees imply a computationally efficient and information-the oretically optimal algorithm for compressed sensing with truncation, such as tha t which may arise from measurement saturation effects. Our result follows from a statistical and computational analysis of the Stochastic Gradient Descent (SGD) algorithm for solving a natural adaption of the LASSO optimization problem that accommodates truncation. This generalizes the works of both: (1) [Daskalakis et al. 2018], where no regularization is needed due to the low dimensionality of t he data, and (2) [Wainright 2009], where the objective function is simple due to the absence of truncation. In order to deal with both truncation and high-dimen sionality at the same time, we develop new techniques that not only generalize t he existing ones but we believe are of independent interest.
**********

Incorporating Pragmatic Reasoning Communication into Emergent Language
Yipeng Kang, Tonghan Wang, Gerard de Melo
Emergentism and pragmatics are two research fields that study the dynamics of li nguistic communication along quite different timescales and intelligence levels. From the perspective of multi-agent reinforcement learning, they correspond to stochastic games with reinforcement training and stage games with opponent aware ness, respectively. Given that their combination has been explored in linguistic s, in this work, we combine computational models of short-term mutual reasoning-based pragmatics with long-term language emergentism. We explore this for agent communication in two settings, referential games and Starcraft II, assessing the relative merits of different kinds of mutual reasoning pragmatics models both e mpirically and theoretically. Our results shed light on their importance for mak ing inroads towards getting more natural, accurate, robust, fine-grained, and su ccinct utterances.
**********

Deep Subspace Clustering with Data Augmentation
Mahdi Abavisani, Alireza Naghizadeh, Dimitris Metaxas, Vishal Patel
The idea behind data augmentation techniques is based on the fact that slight ch anges in the percept do not change the brain cognition. In classification, neura l networks use this fact by applying transformations to the inputs to learn to p redict the same label. However, in deep subspace clustering (DSC), the ground-tr uth labels are not available, and as a result, one cannot easily use data augmen tation techniques. We propose a technique to exploit the benefits of data augmen tation in DSC algorithms. We learn representations that have consistent subspa ces for slightly transformed inputs. In particular, we introduce a temporal ens embling component to the objective function of DSC algorithms to enable the DSC networks to maintain consistent subspaces for random transformations in the inpu t data. In addition, we provide a simple yet effective unsupervised procedure to find efficient data augmentation policies. An augmentation policy is defined as an image processing transformation with a certain magnitude and probability o f being applied to each image in each epoch. We search through the policies in a search space of the most common augmentation policies to find the best policy such that the DSC network yields the highest mean Silhouette coefficient in its

clustering results on a target dataset. Our method achieves state-of-the-art performance on four standard subspace clustering datasets.
**********
An Empirical Process Approach to the Union Bound: Practical Algorithms for Combinatorial and Linear Bandits
Julian Katz-Samuels, Lalit Jain, zohar karnin, Kevin G. Jamieson
This paper proposes near-optimal algorithms for the pure-exploration linear bandit problem in the fixed confidence and fixed budget settings. Leveraging ideas from the theory of suprema of empirical processes, we provide an algorithm whose sample complexity scales with the geometry of the instance and avoids an explicit union bound over the number of arms. Unlike previous approaches which sample based on minimizing a worst-case variance (e.g. G-optimal design), we define an experimental design objective based on the Gaussian-width of the underlying arm set.
We provide a novel lower bound in terms of this objective that highlights its fundamental role in the sample complexity. The sample complexity of our fixed confidence algorithm matches this lower bound, and in addition is computationally efficient for combinatorial classes, e.g. shortest-path, matchings and matroids, where the arm sets can be exponentially large in the dimension. Finally, we propose the first algorithm for linear bandits in the the fixed budget setting. Its guarantee matches our lower bound up to logarithmic factors.
**********
Can Graph Neural Networks Count Substructures?
Zhengdao Chen, Lei Chen, Soledad Villar, Joan Bruna
The ability to detect and count certain substructures in graphs is important for solving many tasks on graph-structured data, especially in the contexts of computational chemistry and biology as well as social network analysis. Inspired by this, we propose to study the expressive power of graph neural networks (GNNs) via their ability to count attributed graph substructures, extending recent works that examine their power in graph isomorphism testing and function approximation. We distinguish between two types of substructure counting: induced-subgraph-count and subgraph-count, and establish both positive and negative answers for popular GNN architectures. Specifically, we prove that Message Passing Neural Networks (MPNNs), 2-Weisfeiler-Lehman (2-WL) and 2-Invariant Graph Networks (2-IGNs) cannot perform induced-subgraph-count of substructures consisting of 3 or more nodes, while they can perform subgraph-count of star-shaped substructures. As an intermediary step, we prove that 2-WL and 2-IGNs are equivalent in distinguishing non-isomorphic graphs, partly answering an open problem raised in Maron et al. (2019). We also prove positive results for k-WL and k-IGNs as well as negative results for k-WL with a finite number of iterations. We then conduct experiments that support the theoretical results for MPNNs and 2-IGNs. Moreover, motivated by substructure counting and inspired by Murphy et al. (2019), we propose the Local Relational Pooling model and demonstrate that it is not only effective for substructure counting but also able to achieve competitive performance on molecular prediction tasks.
**********
A Bayesian Perspective on Training Speed and Model Selection
Clare Lyle, Lisa Schut, Robin Ru, Yarin Gal, Mark van der Wilk
We take a Bayesian perspective to illustrate a connection between training speed and the marginal likelihood in linear models. This provides two major insights: first, that a measure of a model's training speed can be used to estimate its marginal likelihood. Second, that this measure, under certain conditions, predicts the relative weighting of models in linear model combinations trained to minimize a regression loss. We verify our results in model selection tasks for linear models and for the infinite-width limit of deep neural networks. We further provide encouraging empirical evidence that the intuition developed in these settings also holds for deep neural networks trained with stochastic gradient descent. Our results suggest a promising new direction towards explaining why neural networks trained with stochastic gradient descent are biased towards functions that generalize well.

**********
On the Modularity of Hypernetworks
Tomer Galanti, Lior Wolf
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Doubly Robust Off-Policy Value and Gradient Estimation for Deterministic Policies
Nathan Kallus, Masatoshi Uehara
Offline reinforcement learning, wherein one uses off-policy data logged by a fixed behavior policy to evaluate and learn new policies, is crucial in applications where experimentation is limited such as medicine. We study the estimation of policy value and gradient of a deterministic policy from off-policy data when actions are continuous. Targeting deterministic policies, for which action is a deterministic function of state, is crucial since optimal policies are always deterministic (up to ties). In this setting, standard importance sampling and doubly robust estimators for policy value and gradient fail because the density ratio does not exist. To circumvent this issue, we propose several new doubly robust estimators based on different kernelization approaches. We analyze the asymptotic mean-squared error of each of these under mild rate conditions for nuisance estimators. Specifically, we demonstrate how to obtain a rate that is independent of the horizon length.
**********
Provably Efficient Neural GTD for Off-Policy Learning
Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, Mingyi Hong
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Discrete Energy-based Models via Auxiliary-variable Local Exploration
Hanjun Dai, Rishabh Singh, Bo Dai, Charles Sutton, Dale Schuurmans
Discrete structures play an important role in applications
like program language modeling and software engineering.
Current approaches to predicting complex structures typically consider
autoregressive models for their tractability,
with some sacrifice in flexibility.
Energy-based models (EBMs) on the other hand offer a more flexible
and thus more powerful approach to modeling such distributions,
but require partition function estimation.
In this paper we propose \modelshort, a new algorithm for learning conditional
and unconditional EBMs for discrete structured data,
where parameter gradients are estimated using a learned sampler
that mimics local search.
We show that the energy function and sampler can be trained efficiently
via a new variational form of power iteration,
achieving a better trade-off between flexibility and tractability.
Experimentally, we show that learning local search leads to significant
improvements in challenging application domains.
Most notably, we present an energy model guided fuzzer for software testing
that achieves comparable performance
to well engineered fuzzing engines like libfuzzer.
**********
Stable and expressive recurrent vision models
Drew Linsley, Alekh Karkada Ashok, Lakshmi Narasimhan Govindarajan, Rex Liu, Thomas Serre
Primate vision depends on recurrent processing for reliable perception. A growing body of literature also suggests that recurrent connections improve the learni

ng efficiency and generalization of vision models on classic computer vision cha
llenges. Why then, are current large-scale challenges dominated by feedforward n
etworks? We posit that the effectiveness of recurrent vision models is bottlenec
ked by the standard algorithm used for training them, "back-propagation through
time" (BPTT), which has O(N) memory-complexity for training an N step model. Thu
s, recurrent vision model design is bounded by memory constraints, forcing a cho
ice between rivaling the enormous capacity of leading feedforward models or tryi
ng to compensate for this deficit through granular and complex dynamics. Here, w
e develop a new learning algorithm, "contractor recurrent back-propagation" (C-R
BP), which alleviates these issues by achieving constant O(1) memory-complexity
with steps of recurrent processing. We demonstrate that recurrent vision models
trained with C-RBP can detect long-range spatial dependencies in a synthetic con
tour tracing task that BPTT-trained models cannot. We further show that recurren
t vision models trained with C-RBP to solve the large-scale Panoptic Segmentatio
n MS-COCO challenge outperform the leading feedforward approach, with fewer free
parameters. C-RBP is a general-purpose learning algorithm for any application t
hat can benefit from expansive recurrent dynamics. Code and data are available a
t https://github.com/c-rbp.
**********

Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed For
m

Hicham Janati, Boris Muzellec, Gabriel Peyré, Marco Cuturi

Although optimal transport (OT) problems admit closed form solutions in a very f
ew notable cases, e.g. in 1D or between Gaussians, these closed forms have prove
d extremely fecund for practitioners to define tools inspired from the OT geomet
ry. On the other hand, the numerical resolution of OT problems using entropic re
gularization has given rise to many applications, but because there are no known
closed-form solutions for entropic regularized OT problems, these approaches ar
e mostly algorithmic, not informed by elegant closed forms. In this paper, we pr
opose to fill the void at the intersection between these two schools of thought
in OT by proving that the entropy-regularized optimal transport problem between
two Gaussian measures admits a closed form. Contrary to the unregularized case,
for which the explicit form is given by the Wasserstein-Bures distance, the clos
ed form we obtain is differentiable everywhere, even for Gaussians with degenera
te covariance matrices. We obtain this closed form solution by solving the fixed
-point equation behind Sinkhorn's algorithm, the default method for computing en
tropic regularized OT. Remarkably, this approach extends to the generalized unba
lanced case --- where Gaussian measures are scaled by positive constants. This e
xtension leads to a closed form expression for unbalanced Gaussians as well, and
highlights the mass transportation / destruction trade-off seen in unbalanced o
ptimal transport. Moreover, in both settings, we show that the optimal transport
ation plans are (scaled) Gaussians and provide analytical formulas of their para
meters. These formulas constitute the first non-trivial closed forms for entropy
-regularized optimal transport, thus providing a ground truth for the analysis o
f entropic OT and Sinkhorn's algorithm.
**********

BRP-NAS: Prediction-based NAS using GCNs

Lukasz Dudziak, Thomas Chau, Mohamed Abdelfattah, Royson Lee, Hyeji Kim, Nichola
s Lane

Neural architecture search (NAS) enables researchers to automatically explore br
oad design spaces in order to improve efficiency of neural networks. This effici
ency is especially important in the case of on-device deployment, where improvem
ents in accuracy should be balanced out with computational demands of a model. I
n practice, performance metrics of model are computationally expensive to obtain
. Previous work uses a proxy (e.g., number of operations) or a layer-wise measur
ement of neural network layers to estimate end-to-end hardware performance but t
he imprecise prediction diminishes the quality of NAS. To address this problem,
we propose BRP-NAS, an efficient hardware-aware NAS enabled by an accurate perfo
rmance predictor-based on graph convolutional network (GCN). What is more, we in
vestigate prediction quality on different metrics and show that sample efficienc

y of the predictor-based NAS can be improved by considering binary relations of models and an iterative data selection strategy. We show that our proposed method outperforms all prior methods on NAS-Bench-101, NAS-Bench-201 and DARTS. Finally, to raise awareness of the fact that accurate latency estimation is not a trivial task, we release LatBench -- a latency dataset of NAS-Bench-201 models running on a broad range of devices.
**********

Deep Shells: Unsupervised Shape Correspondence with Optimal Transport
Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, Daniel Cremers
We propose a novel unsupervised learning approach to 3D shape correspondence that builds a multiscale matching pipeline into a deep neural network. This approach is based on smooth shells, the current state-of-the-art axiomatic correspondence method, which requires an a priori stochastic search over the space of initial poses. Our goal is to replace this costly preprocessing step by directly learning good initializations from the input surfaces. To that end, we systematically derive a fully differentiable, hierarchical matching pipeline from entropy regularized optimal transport. This allows us to combine it with a local feature extractor based on smooth, truncated spectral convolution filters. Finally, we show that the proposed unsupervised method significantly improves over the state-of-the-art on multiple datasets, even in comparison to the most recent supervised methods. Moreover, we demonstrate compelling generalization results by applying our learned filters to examples that significantly deviate from the training set.
**********

ISTA-NAS: Efficient and Consistent Neural Architecture Search by Sparse Coding
Yibo Yang, Hongyang Li, Shan You, Fei Wang, Chen Qian, Zhouchen Lin
Neural architecture search (NAS) aims to produce the optimal sparse solution from a high-dimensional space spanned by all candidate connections. Current gradient-based NAS methods commonly ignore the constraint of sparsity in the search phase, but project the optimized solution onto a sparse one by post-processing. As a result, the dense super-net for search is inefficient to train and has a gap with the projected architecture for evaluation. In this paper, we formulate neural architecture search as a sparse coding problem. We perform the differentiable search on a compressed lower-dimensional space that has the same validation loss as the original sparse solution space, and recover an architecture by solving the sparse coding problem. The differentiable search and architecture recovery are optimized in an alternate manner. By doing so, our network for search at each update satisfies the sparsity constraint and is efficient to train. In order to also eliminate the depth and width gap between the network in search and the target-net in evaluation, we further propose a method to search and evaluate in one stage under the target-net settings. When training finishes, architecture variables are absorbed into network weights. Thus we get the searched architecture and optimized parameters in a single run. In experiments, our two-stage method on CIFAR-10 requires only 0.05 GPU-day for search. Our one-stage method produces state-of-the-art performances on both CIFAR-10 and ImageNet at the cost of only evaluation time.
**********

Rel3D: A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D
Ankit Goyal, Kaiyu Yang, Dawei Yang, Jia Deng
Understanding spatial relations (e.g., laptop on table) in visual input is important for both humans and robots. Existing datasets are insufficient as they lack large-scale, high-quality 3D ground truth information, which is critical for learning spatial relations. In this paper, we fill this gap by constructing Rel3D: the first large-scale, human-annotated dataset for grounding spatial relations in 3D. Rel3D enables quantifying the effectiveness of 3D information in predicting spatial relations on large-scale human data. Moreover, we propose minimally contrastive data collection---a novel crowdsourcing method for reducing dataset bias. The 3D scenes in our dataset come in minimally contrastive pairs: two scenes in a pair are almost identical, but a spatial relation holds in one and fails in the other. We empirically validate that minimally contrastive examples can diagnose issues with current relation detection models as well as lead to sample-e

fficient training. Code and data are available at https://github.com/princeton-vl/Rel3D.
**********

Regularizing Black-box Models for Improved Interpretability
Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, Ameet Talwalkar
Most of the work on interpretable machine learning has focused on designing either inherently interpretable models, which typically trade-off accuracy for interpretability, or post-hoc explanation systems, whose explanation quality can be unpredictable. Our method, ExpO, is a hybridization of these approaches that regularizes a model for explanation quality at training time. Importantly, these regularizers are differentiable, model agnostic, and require no domain knowledge to define. We demonstrate that post-hoc explanations for ExpO-regularized models have better explanation quality, as measured by the common fidelity and stability metrics. We verify that improving these metrics leads to significantly more useful explanations with a user study on a realistic task.
**********

Trust the Model When It Is Confident: Masked Model-based Actor-Critic
Feiyang Pan, Jia He, Dandan Tu, Qing He
It is a popular belief that model-based Reinforcement Learning (RL) is more sample efficient than model-free RL, but in practice, it is not always true due to overweighed model errors. In complex and noisy settings, model-based RL tends to have trouble using the model if it does not know when to trust the model.
**********

Semi-Supervised Neural Architecture Search
Renqian Luo, Xu Tan, Rui Wang, Tao Qin, Enhong Chen, Tie-Yan Liu
Neural architecture search (NAS) relies on a good controller to generate better architectures or predict the accuracy of given architectures. However, training the controller requires both abundant and high-quality pairs of architectures and their accuracy, while it is costly to evaluate an architecture and obtain its accuracy. In this paper, we propose SemiNAS, a semi-supervised NAS approach that leverages numerous unlabeled architectures (without evaluation and thus nearly no cost). Specifically, SemiNAS 1) trains an initial accuracy predictor with a small set of architecture-accuracy data pairs; 2) uses the trained accuracy predictor to predict the accuracy of large amount of architectures (without evaluation); and 3) adds the generated data pairs to the original data to further improve the predictor. The trained accuracy predictor can be applied to various NAS algorithms by predicting the accuracy of candidate architectures for them. SemiNAS has two advantages: 1) It reduces the computational cost under the same accuracy guarantee. On NASBench-101 benchmark dataset, it achieves comparable accuracy with gradient-based method while using only 1/7 architecture-accuracy pairs. 2) It achieves higher accuracy under the same computational cost. It achieves 94.02% test accuracy on NASBench-101, outperforming all the baselines when using the same number of architectures. On ImageNet, it achieves 23.5% top-1 error rate (under 600M FLOPS constraint) using 4 GPU-days for search. We further apply it to LJSpeech text to speech task and it achieves 97% intelligibility rate in the low-resource setting and 15% test error rate in the robustness setting, with 9%, 7% improvements over the baseline respectively.
**********

Consistency Regularization for Certified Robustness of Smoothed Classifiers
Jongheon Jeong, Jinwoo Shin
A recent technique of randomized smoothing has shown that the worst-case (adversarial) l2-robustness can be transformed into the average-case Gaussian-robustness by "smoothing" a classifier, i.e., by considering the averaged prediction over Gaussian noise. In this paradigm, one should rethink the notion of adversarial robustness in terms of generalization ability of a classifier under noisy observations. We found that the trade-off between accuracy and certified robustness of smoothed classifiers can be greatly controlled by simply regularizing the prediction consistency over noise. This relationship allows us to design a robust training objective without approximating a non-existing smoothed classifier, e.g.,

via soft smoothing. Our experiments under various deep neural network architectures and datasets show that the "certified" l2-robustness can be dramatically improved with the proposed regularization, even achieving better or comparable results to the state-of-the-art approaches with significantly less training costs and hyperparameters.
**********

Robust Multi-Agent Reinforcement Learning with Model Uncertainty
Kaiqing Zhang, TAO SUN, Yunzhe Tao, Sahika Genc, Sunil Mallya, Tamer Basar
In this work, we study the problem of multi-agent reinforcement learning (MARL) with model uncertainty, which is referred to as robust MARL. This is naturally motivated by some multi-agent applications where each agent may not have perfectly accurate knowledge of the model, e.g., all the reward functions of other agents. Little a priori work on MARL has accounted for such uncertainties, neither in problem formulation nor in algorithm design. In contrast, we model the problem as a robust Markov game, where the goal of all agents is to find policies such that no agent has the incentive to deviate, i.e., reach some equilibrium point, which is also robust to the possible uncertainty of the MARL model. We first introduce the solution concept of robust Nash equilibrium in our setting, and develop a Q-learning algorithm to find such equilibrium policies, with convergence guarantees under certain conditions. In order to handle possibly enormous state-action spaces in practice, we then derive the policy gradients for robust MARL, and develop an actor-critic algorithm with function approximation. Our experiments demonstrate that the proposed algorithm outperforms several baseline MARL methods that do not account for the model uncertainty, in several standard but uncertain cooperative and competitive MARL environments.
**********

SIRI: Spatial Relation Induced Network For Spatial Description Resolution
peiyao wang, Weixin Luo, Yanyu Xu, Haojie Li, Shugong Xu, Jianyu Yang, Shenghua Gao
Spatial Description Resolution, as a language-guided localization task, is proposed for target location in a panoramic street view, given corresponding language descriptions. Explicitly characterizing an object-level relationship while distilling spatial relationships are currently absent but crucial to this task. Mimicking humans, who sequentially traverse spatial relationship words and objects with a first-person view to locate their target, we propose a novel spatial relationship induced (SIRI) network. Specifically, visual features are firstly correlated at an implicit object-level in a projected latent space; then they are distilled by each spatial relationship word, resulting in each differently activated feature representing each spatial relationship. Further, we introduce global position priors to fix the absence of positional information, which may result in global positional reasoning ambiguities. Both the linguistic and visual features are concatenated to finalize the target localization. Experimental results on the Touchdown show that our method is around 24\% better than the state-of-the-art method in terms of accuracy, measured by an 80-pixel radius. Our method also generalizes well on our proposed extended dataset collected using the same settings as Touchdown. The code for this project is
publicly available at https://github.com/wong-puiyiu/siri-sdr.
**********

Adaptive Shrinkage Estimation for Streaming Graphs
Nesreen Ahmed, Nick Duffield
Networks are a natural representation of complex systems across the sciences, and higher-order dependencies are central to the understanding and modeling of these systems. However, in many practical applications such as online social networks, networks are massive, dynamic, and naturally streaming, where pairwise interactions among vertices become available one at a time in some arbitrary order. The massive size and streaming nature of these networks allow only partial observation, since it is infeasible to analyze the entire network. Under such scenarios, it is challenging to study the higher-order structural and connectivity patterns of streaming networks. In this work, we consider the fundamental problem of estimating the higher-order dependencies using adaptive sampling. We propose a n

ovel adaptive, single-pass sampling framework and unbiased estimators for higher
-order network analysis of large streaming networks. Our algorithms exploit adap
tive techniques to identify edges that are highly informative for efficiently es
timating the higher-order structure of streaming networks from small sample data
. We also introduce a novel James-Stein shrinkage estimator to reduce the estima
tion error. Our approach is fully analytic, computationally efficient, and can b
e incrementally updated in a streaming setting. Numerical experiments on large n
etworks show that our approach is superior to baseline methods.
**********

Make One-Shot Video Object Segmentation Efficient Again
Tim Meinhardt, Laura Leal-Taixé
Video object segmentation (VOS) describes the task of segmenting a set of object
s in each frame of a video. In the semi-supervised setting, the first mask of ea
ch object is provided at test time. Following the one-shot principle, fine-tunin
g VOS methods train a segmentation model separately on each given object mask. H
owever, recently the VOS community has deemed such a test time optimization and
its impact on the test runtime as unfeasible. To mitigate the inefficiencies of
previous fine-tuning approaches, we present efficient One-Shot Video Object Segm
entation (e-OSVOS). In contrast to most VOS approaches, e-OSVOS decouples the ob
ject detection task and predicts only local segmentation masks by applying a mod
ified version of Mask R-CNN. The one-shot test runtime and performance are optim
ized without a laborious and handcrafted hyperparameter search. To this end, we
meta learn the model initialization and learning rates for the test time optimiz
ation. To achieve an optimal learning behavior, we predict individual learning r
ates at a neuron level. % a pair of learning rates for the weights tensor and sc
alar bias of each neuron. Furthermore, we apply an online adaptation to address
the common performance degradation throughout a sequence by continuously fine-tu
ning the model on previous mask predictions supported by a frame-to-frame boundi
ng box propagation. % through changing online appearance -> online adaptation fo
r free. bounding box propagation. e-OSVOS provides state-of-the-art results on D
AVIS 2016, DAVIS 2017 and YouTube-VOS for one-shot fine-tuning methods while red
ucing the test runtime substantially.
**********

Depth Uncertainty in Neural Networks
Javier Antoran, James Allingham, José Miguel Hernández-Lobato
Existing methods for estimating uncertainty in deep learning tend to require mul
tiple forward passes, making them unsuitable for applications where computationa
l resources are limited. To solve this, we perform probabilistic reasoning over
the depth of neural networks. Different depths correspond to subnetworks which s
hare weights and whose predictions are combined via marginalisation, yielding mo
del uncertainty. By exploiting the sequential structure of feed-forward networks
, we are able to both evaluate our training objective and make predictions with
a single forward pass. We validate our approach on real-world regression and ima
ge classification tasks. Our approach provides uncertainty calibration, robustne
ss to dataset shift, and accuracies competitive with more computationally expens
ive baselines.
**********

Non-Euclidean Universal Approximation
Anastasis Kratsios, Ievgen Bilokopytov
Modifications to a neural network's input and output layers are often required t
o accommodate the specificities of most practical learning tasks. However, the i
mpact of such changes on architecture's approximation capabilities is largely no
t understood. We present general conditions describing feature and readout maps
that preserve an architecture's ability to approximate any continuous functions
uniformly on compacts. As an application, we show that if an architecture is cap
able of universal approximation, then modifying its final layer to produce binar
y values creates a new architecture capable of deterministically approximating a
ny classifier. In particular, we obtain guarantees for deep CNNs, deep ffNN, and
 universal Gaussian processes.  Our results also have consequences within the sc
ope of geometric deep learning. Specifically, when the input and output spaces a

re Hadamard manifolds, we obtain geometrically meaningful feature and readout maps satisfying our criteria. Consequently, commonly used non-Euclidean regression models between spaces of symmetric positive definite matrices are extended to universal DNNs. The same result allows us to show that the hyperbolic feed-forward networks, used for hierarchical learning, are universal. Our result is also used to show that the common practice of randomizing all but the last two layers of a DNN produces a universal family of functions with probability one. We also provide conditions on a DNN's first (resp. last) few layer's connections and activation function which guarantee that these layer's can have a width equal to the input (resp. output) space's dimension while not negatively effecting the architecture's approximation capabilities.
**********

Constraining Variational Inference with Geometric Jensen-Shannon Divergence
Jacob Deasy, Nikola Simidjievski, Pietro Lió
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Gibbs Sampling with People
Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, Nori Jacoby
A core problem in cognitive science and machine learning is to understand how humans derive semantic representations from perceptual objects, such as color from an apple, pleasantness from a musical chord, or seriousness from a face. Markov Chain Monte Carlo with People (MCMCP) is a prominent method for studying such representations, in which participants are presented with binary choice trials constructed such that the decisions follow a Markov Chain Monte Carlo acceptance rule. However, while MCMCP has strong asymptotic properties, its binary choice paradigm generates relatively little information per trial, and its local proposal function makes it slow to explore the parameter space and find the modes of the distribution. Here we therefore generalize MCMCP to a continuous-sampling paradigm, where in each iteration the participant uses a slider to continuously manipulate a single stimulus dimension to optimize a given criterion such as ‘pleasantness’. We formulate both methods from a utility-theory perspective, and show that the new method can be interpreted as ‘Gibbs Sampling with People’ (GSP). Further, we introduce an aggregation parameter to the transition step, and show that this parameter can be manipulated to flexibly shift between Gibbs sampling and deterministic optimization. In an initial study, we show GSP clearly outperforming MCMCP; we then show that GSP provides novel and interpretable results in three other domains, namely musical chords, vocal emotions, and faces. We validate these results through large-scale perceptual rating experiments. The final experiments use GSP to navigate the latent space of a state-of-the-art image synthesis network (StyleGAN), a promising approach for applying GSP to high-dimensional perceptual spaces. We conclude by discussing future cognitive applications and ethical implications.
**********

HM-ANN: Efficient Billion-Point Nearest Neighbor Search on Heterogeneous Memory
Jie Ren, Minjia Zhang, Dong Li
The state-of-the-art approximate nearest neighbor search (ANNS) algorithms face a fundamental tradeoff between query latency and accuracy, because of small main memory capacity: To store indices in main memory for short query latency, the ANNS algorithms have to limit dataset size or use a quantization scheme which hurts search accuracy. The emergence of heterogeneous memory (HM) brings a solution to significantly increase memory capacity and break the above tradeoff: Using HM, billions of data points can be placed in the main memory on a single machine without using any data compression. However, HM consists of both fast (but small) memory and slow (but large) memory, and using HM inappropriately slows down query significantly.
In this work, we present a novel graph-based similarity search algorithm called

HM-ANN, which takes both memory and data heterogeneity into consideration and en
ables billion-scale similarity search on a single node without using compression
. On two billion-sized datasets BIGANN and DEEP1B, HM-ANN outperforms state-of-t
he-art compression-based solutions such as L&C and IMI+OPQ in recall-vs-latency
by a large margin, obtaining 46% higher recall under the same search latency. We
 also extend existing graph-based methods such as HNSW and NSG with two strong b
aseline implementations on HM. At billion-point scale, HM-ANN is 2X and 5.8X fas
ter than our HNSWand NSG baselines respectively to reach the same accuracy.
**********

FrugalML: How to use ML Prediction APIs more accurately and cheaply
Lingjiao Chen, Matei Zaharia, James Y. Zou
Offering prediction APIs for fee is a fast growing industry and is an important
aspect of machine learning as a service. While many such services are available,
 the heterogeneity in their price and performance makes it challenging for users
 to decide which API or combination of APIs to use for their own data and budget
. We take a first step towards addressing this challenge by proposing FrugalML,
a principled framework that jointly learns the strength and weakness of each API
 on different data, and performs an efficient optimization to automatically iden
tify the best sequential strategy to adaptively use the available APIs within a
budget constraint. Our theoretical analysis shows that natural sparsity in the f
ormulation can be leveraged to make FrugalML efficient. We conduct systematic ex
periments using ML APIs from Google, Microsoft, Amazon, IBM, Baidu and other pro
viders for tasks including facial emotion recognition, sentiment analysis and sp
eech recognition. Across various tasks, FrugalML can achieve up to 90% cost redu
ction while matching the accuracy of the best single API, or up to 5% better acc
uracy while matching the best API’s cost.
**********

Sharp Representation Theorems for ReLU Networks with Precise Dependence on Depth
Guy Bresler, Dheeraj Nagaraj
We prove dimension free representation results for neural networks with D ReLU l
ayers under square loss for a class of functions G_D defined in the paper. These
 results capture the precise benefits of depth in the following sense:
**********

Shared Experience Actor-Critic for Multi-Agent Reinforcement Learning
Filippos Christianos, Lukas Schäfer, Stefano Albrecht
Exploration in multi-agent reinforcement learning is a challenging problem, espe
cially in environments with sparse rewards. We propose a general method for effi
cient exploration by sharing experience amongst agents.  Our proposed algorithm,
 called shared Experience Actor-Critic(SEAC), applies experience sharing in an a
ctor-critic framework by combining the gradients of different agents. We evaluat
e SEAC in a collection of sparse-reward multi-agent environments and find that i
t consistently outperforms several baselines and state-of-the-art algorithms by
learning in fewer steps and converging to higher returns. In some harder environ
ments, experience sharing makes the difference between learning to solve the tas
k and not learning at all.
**********

Monotone operator equilibrium networks
Ezra Winston, J. Zico Kolter
Implicit-depth models such as Deep Equilibrium Networks have recently been shown
 to match or exceed the performance of traditional deep networks while being muc
h more memory efficient. However, these models suffer from unstable convergence
to a solution and lack guarantees that a solution exists. On the other hand, Neu
ral ODEs, another class of implicit-depth models, do guarantee existence of a un
ique solution but perform poorly compared with traditional networks. In this pap
er, we develop a new class of implicit-depth model based on the theory of monoto
ne operators, the Monotone Operator Equilibrium Network (monDEQ). We show the cl
ose connection between finding the equilibrium point of an implicit network and
solving a form of monotone operator splitting problem, which admits efficient so
lvers with guaranteed, stable convergence. We then develop a parameterization of
 the network which ensures that all operators remain monotone, which guarantees

the existence of a unique equilibrium point. Finally, we show how to instantiate several versions of these models, and implement the resulting iterative solvers, for structured linear operators such as multi-scale convolutions. The resulting models vastly outperform the Neural ODE-based models while also being more computationally efficient. Code is available at http://github.com/locuslab/monotone_opnet.

**********

When and How to Lift the Lockdown? Global COVID-19 Scenario Analysis and Policy Assessment using Compartmental Gaussian Processes

Zhaozhi Qian, Ahmed M. Alaa, Mihaela van der Schaar

The coronavirus disease 2019 (COVID-19) global pandemic has led many countries to impose unprecedented lockdown measures in order to slow down the outbreak. Questions on whether governments have acted promptly enough, and whether lockdown measures can be lifted soon have since been central in public discourse. Data-driven models that predict COVID-19 fatalities under different lockdown policy scenarios are essential for addressing these questions, and for informing governments on future policy directions. To this end, this paper develops a Bayesian model for predicting the effects of COVID-19 containment policies in a global context — we treat each country as a distinct data point, and exploit variations of policies across countries to learn country-specific policy effects. Our model utilizes a two-layer Gaussian process (GP) prior — the lower layer uses a compartmental SEIR (Susceptible, Exposed, Infected, Recovered) model as a prior mean function with “country-and-policy-specific” parameters that capture fatality curves under different “counterfactual” policies within each country, whereas the upper layer is shared across all countries, and learns lower-layer SEIR parameters as a function of country features and policy indicators. Our model combines the solid mechanistic foundations of SEIR models (Bayesian priors) with the flexible data-driven modeling and gradient-based optimization routines of machine learning (Bayesian posteriors) — i.e., the entire model is trained end-to-end via stochastic variational inference. We compare the projections of our model with other models listed by the Center for Disease Control (CDC), and provide scenario analyses for various lockdown and reopening strategies highlighting their impact on COVID-19 fatalities.

**********

Unsupervised Learning of Lagrangian Dynamics from Images for Prediction and Control

Yaofeng Desmond Zhong, Naomi Leonard

Recent approaches for modelling dynamics of physical systems with neural networks enforce Lagrangian or Hamiltonian structure to improve prediction and generalization. However, when coordinates are embedded in high-dimensional data such as images, these approaches either lose interpretability or can only be applied to one particular example. We introduce a new unsupervised neural network model that learns Lagrangian dynamics from images, with interpretability that benefits prediction and control. The model infers Lagrangian dynamics on generalized coordinates that are simultaneously learned with a coordinate-aware variational autoencoder (VAE). The VAE is designed to account for the geometry of physical systems composed of multiple rigid bodies in the plane. By inferring interpretable Lagrangian dynamics, the model learns physical system properties, such as kinetic and potential energy, which enables long-term prediction of dynamics in the image space and synthesis of energy-based controllers.

**********

High-Dimensional Sparse Linear Bandits

Botao Hao, Tor Lattimore, Mengdi Wang

Stochastic linear bandits with high-dimensional sparse features are a practical model for a variety of domains, such as personalized medicine and online advertising. We derive a novel $O(n^{2/3})$ dimension-free minimax regret lower bound for sparse linear bandits in the data-poor regime where the horizon is larger than the ambient dimension and where the feature vectors admit a well-conditioned exploration distribution. This is complemented by a nearly matching upper bound for an explore-then-commit algorithm showing that that $O(n^{2/3})$ is the optimal ra

te in the data-poor regime. The results complement existing bounds for the data-rich regime and also provide another example where carefully balancing the trade-off between information and regret is necessary. Finally, we prove a dimension-free O(\sqrt{n}) regret upper bound under an additional assumption on the magnitude of the signal for relevant features.
**********

Non-Stochastic Control with Bandit Feedback
Paula Gradu, John Hallman, Elad Hazan
We study the problem of controlling a linear dynamical system with adversarial perturbations where the only feedback available to the controller is the scalar loss, and the loss function itself is unknown. For this problem, with either a known or unknown system, we give an efficient sublinear regret algorithm. The main algorithmic difficulty is the dependence of the loss on past controls. To overcome this issue, we propose an efficient algorithm for the general setting of bandit convex optimization for loss functions with memory, which may be of independent interest.
**********

Generalized Leverage Score Sampling for Neural Networks
Jason D. Lee, Ruoqi Shen, Zhao Song, Mengdi Wang, zheng Yu
Leverage score sampling is a powerful technique that originates from theoretical computer science, which can be used to speed up a large number of fundamental questions, e.g. linear regression, linear programming, semi-definite programming, cutting plane method, graph sparsification, maximum matching and max-flow. Recently, it has been shown that leverage score sampling helps to accelerate kernel methods [Avron, Kapralov, Musco, Musco, Velingker and Zandieh 17]. In this work, we generalize the results in [Avron, Kapralov, Musco, Musco, Velingker and Zandieh 17] to a broader class of kernels. We further bring the leverage score sampling into the field of deep learning theory.
1. We show the connection between the initialization for neural network training and approximating the neural tangent kernel with random features.
2. We prove the equivalence between regularized neural network and neural tangent kernel ridge regression under the initialization of both classical random Gaussian and leverage score sampling.
**********

An Optimal Elimination Algorithm for Learning a Best Arm
Avinatan Hassidim, Ron Kupfer, Yaron Singer
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  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 Projection-free Algorithms for Saddle Point Problems
Cheng Chen, Luo Luo, Weinan Zhang, Yong 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.
**********

A mathematical model for automatic differentiation in machine learning
Jérôme Bolte, Edouard Pauwels
Automatic differentiation, as implemented today, does not have a simple mathematical model adapted to the needs of modern machine learning. In this work we articulate the relationships between differentiation of programs as implemented in practice, and differentiation of nonsmooth functions. To this end we provide a simple class of functions, a nonsmooth calculus, and show how they apply to stochastic approximation methods. We also evidence the issue of artificial critical points created by algorithmic differentiation and show how usual methods avoid these points with probability one.
**********

Unsupervised Text Generation by Learning from Search

Jingjing Li, Zichao Li, Lili Mou, Xin Jiang, Michael Lyu, Irwin King
In this work, we propose TGLS, a novel framework for unsupervised Text Generatio n by Learning from Search. We start by applying a strong search algorithm (in pa rticular, simulated annealing) towards a heuristically defined objective that (r oughly) estimates the quality of sentences. Then, a conditional generative model learns from the search results, and meanwhile smooth out the noise of search. T he alternation between search and learning can be repeated for performance boots trapping. We demonstrate the effectiveness of TGLS on two real-world natural lan guage generation tasks, unsupervised paraphrasing and text formalization. Our mo del significantly outperforms unsupervised baseline methods in both tasks. Espec ially, it achieves comparable performance to strong supervised methods for parap hrase generation.
**********
Learning Compositional Rules via Neural Program Synthesis
Maxwell Nye, Armando Solar-Lezama, Josh Tenenbaum, Brenden M. Lake
Many aspects of human reasoning, including language, require learning rules from very little data. Humans can do this, often learning systematic rules from very few examples, and combining these rules to form compositional rule-based system s. Current neural architectures, on the other hand, often fail to generalize in a compositional manner, especially when evaluated in ways that vary systematical ly from training. In this work, we present a neuro-symbolic model which learns e ntire rule systems from a small set of examples. Instead of directly predicting outputs from inputs, we train our model to induce the explicit system of rules g overning a set of previously seen examples, drawing upon techniques from the neu ral program synthesis literature. Our rule-synthesis approach outperforms neural meta-learning techniques in three domains: an artificial instruction-learning d omain used to evaluate human learning, the SCAN challenge datasets, and learning rule-based translations of number words into integers for a wide range of human languages.
**********
Incorporating BERT into Parallel Sequence Decoding with Adapters
Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, 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-auth ors prior to requesting a name change in the electronic proceedings.
**********
Estimating Fluctuations in Neural Representations of Uncertain Environments
Sahand Farhoodi, Mark Plitt, Lisa Giocomo, Uri Eden
Neural Coding analyses often reflect an assumption that neural populations respo nd uniquely and consistently to particular stimuli. For example, analyses of spa tial remapping in hippocampal populations often assume that each environment has one unique representation and that remapping occurs over long time scales as an animal traverses between distinct environments. However, as neuroscience experi ments begin to explore more naturalistic tasks and stimuli, and reflect more amb iguity in neural representations, methods for analyzing population neural codes must adapt to reflect these features. In this paper, we develop a new state-spac e modeling framework to address two important issues related to remapping. First , neurons may exhibit significant trial-to-trial or moment-to-moment variability in the firing patterns used to represent a particular environment or stimulus. Second, in ambiguous environments and tasks that involve cognitive uncertainty, neural populations may rapidly fluctuate between multiple representations. The s tate-space model addresses these two issues by integrating an observation model, which allows for multiple representations of the same stimulus or environment, with a state model, which characterizes the moment-by-moment probability of a sh ift in the neural representation. These models allow us to compute instantaneous estimates of the stimulus or environment currently represented by the populatio n. We demonstrate the application of this approach to the analysis of population activity in the CA1 region of hippocampus of a mouse moving through ambiguous v irtual environments. Our analyses demonstrate that many hippocampal cells expres

s significant trial-to-trial variability in their representations and that the p
opulation representation can fluctuate rapidly between environments within a sin
gle trial when spatial cues are most ambiguous.
**********
Discover, Hallucinate, and Adapt: Open Compound Domain Adaptation for Semantic S
egmentation
KwanYong Park, Sanghyun Woo, Inkyu Shin, In So Kweon
Unsupervised domain adaptation (UDA) for semantic segmentation has been attracti
ng attention recently, as it could be beneficial for various label-scarce real-w
orld scenarios (e.g., robot control, autonomous driving, medical imaging, etc.).
 Despite the significant progress in this field, current works mainly focus on a
 single-source single-target setting, which cannot handle more practical setting
s of multiple targets or even unseen targets.
In this paper, we investigate open compound domain adaptation (OCDA), which deal
s with mixed and novel situations at the same time, for semantic segmentation.
We present a novel framework based on three main design principles: discover, ha
llucinate, and adapt. The scheme first clusters compound target data based on st
yle, discovering multiple latent domains (discover). Then, it hallucinates multi
ple latent target domains in source by using image-translation (hallucinate). Th
is step ensures the latent domains in the source and the target to be paired. Fi
nally, target-to-source alignment is learned separately between domains (adapt).
 In high-level, our solution replaces a hard OCDA problem with much easier multi
ple UDA problems.
We evaluate our solution on standard benchmark GTA to C-driving, and achieved ne
w state-of-the-art results.
**********
SURF: A Simple, Universal, Robust, Fast Distribution Learning Algorithm
Yi Hao, Ayush Jain, Alon Orlitsky, Vaishakh Ravindrakumar
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Understanding Approximate Fisher Information for Fast Convergence of Natural Gra
dient Descent in Wide Neural Networks
Ryo Karakida, Kazuki Osawa
Natural Gradient Descent (NGD) helps to accelerate the convergence of gradient d
escent dynamics, but it requires approximations in large-scale deep neural netwo
rks because of its high computational cost. Empirical studies have confirmed tha
t some NGD methods with approximate Fisher information converge sufficiently fas
t in practice. Nevertheless, it remains unclear from the theoretical perspective
 why and under what conditions such heuristic approximations work well. In this
work, we reveal that, under specific conditions, NGD with approximate Fisher inf
ormation achieves the same fast convergence to global minima as exact NGD. We co
nsider deep neural networks in the infinite-width limit, and analyze the asympto
tic training dynamics of NGD in function space via the neural tangent kernel. In
 the function space, the training dynamics with the approximate Fisher informati
on are identical to those with the exact Fisher information, and they converge q
uickly. The fast convergence holds in layer-wise approximations; for instance, i
n block diagonal approximation where each block corresponds to a layer as well a
s in block tri-diagonal and K-FAC approximations. We also find that a unit-wise
approximation achieves the same fast convergence under some assumptions. All of
these different approximations have an isotropic gradient in the function space,
 and this plays a fundamental role in achieving the same convergence properties
in training. Thus, the current study gives a novel and unified theoretical found
ation with which to understand NGD methods in deep learning.
**********
General Transportability of Soft Interventions: Completeness Results
Juan Correa, Elias Bareinboim
Requests for name changes in the electronic proceedings will be accepted with no

questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

GAIT-prop: A biologically plausible learning rule derived from backpropagation of error
Nasir Ahmad, Marcel A. J. van Gerven, Luca Ambrogioni
Traditional backpropagation of error, though a highly successful algorithm for learning in artificial neural network models, includes features which are biologically implausible for learning in real neural circuits. An alternative called target propagation proposes to solve this implausibility by using a top-down model of neural activity to convert an error at the output of a neural network into layer-wise and plausible ‘targets’ for every unit. These targets can then be used to produce weight updates for network training. However, thus far, target propagation has been heuristically proposed without demonstrable equivalence to backpropagation. Here, we derive an exact correspondence between backpropagation and a modified form of target propagation (GAIT-prop) where the target is a small perturbation of the forward pass. Specifically, backpropagation and GAIT-prop give identical updates when synaptic weight matrices are orthogonal. In a series of simple computer vision experiments, we show near-identical performance between backpropagation and GAIT-prop with a soft orthogonality-inducing regularizer.
**********

Lipschitz Bounds and Provably Robust Training by Laplacian Smoothing
Vishaal Krishnan, Abed AlRahman Al Makdah, Fabio Pasqualetti
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

SCOP: Scientific Control for Reliable Neural Network Pruning
Yehui Tang, Yunhe Wang, Yixing Xu, Dacheng Tao, Chunjing XU, Chao Xu, Chang Xu
This paper proposes a reliable neural network pruning algorithm by setting up a scientific control. Existing pruning methods have developed various hypotheses to approximate the importance of filters to the network and then execute filter pruning accordingly. To increase the reliability of the results, we prefer to have a more rigorous research design by including a scientific control group as an essential part to minimize the effect of all factors except the association between the filter and expected network output. Acting as a control group, knockoff feature is generated to mimic the feature map produced by the network filter, but they are conditionally independent of the example label given the real feature map. We theoretically suggest that the knockoff condition can be approximately preserved given the information propagation of network layers.  Besides the real feature map on an intermediate layer, the corresponding knockoff feature is brought in as another auxiliary input signal for the subsequent layers.
Redundant filters can be discovered in the adversarial process of different features. Through experiments, we demonstrate the superiority of the proposed algorithm over state-of-the-art methods. For example, our method can reduce 57.8% parameters and 60.2% FLOPs of ResNet-101 with only 0.01% top-1 accuracy loss on ImageNet.
**********

Provably Consistent Partial-Label Learning
Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, Masashi Sugiyama
Partial-label learning (PLL) is a multi-class classification problem, where each training example is associated with a set of candidate labels. Even though many practical PLL methods have been proposed in the last two decades, there lacks a theoretical understanding of the consistency of those methods - none of the PLL methods hitherto possesses a generation process of candidate label sets, and then it is still unclear why such a method works on a specific dataset and when it may fail given a different dataset. In this paper, we propose the first generation model of candidate label sets, and develop two PLL methods that are guarante

ed to be provably consistent, i.e., one is risk-consistent and the other is clas
sifier-consistent. Our methods are advantageous, since they are compatible with
any deep network or stochastic optimizer. Furthermore, thanks to the generation
model, we would be able to answer the two questions above by testing if the gene
ration model matches given candidate label sets. Experiments on benchmark and re
al-world datasets validate the effectiveness of the proposed generation model an
d two PLL methods.
**********

Robust, Accurate Stochastic Optimization for Variational Inference
Akash Kumar Dhaka, Alejandro Catalina, Michael R. Andersen, Måns Magnusson, Jona
than Huggins, Aki Vehtari
We examine the accuracy of black box variational posterior approximations for pa
rametric models in a probabilistic programming context. The performance of these
approximations depends on (1) how well the variational family approximates the
true posterior distribution, (2) the choice of divergence, and (3) the optimizat
ion of the variational objective. We show that even when the true variational fa
mily is used, high-dimensional posteriors can be very poorly approximated using
common stochastic gradient descent (SGD) optimizers. Motivated by recent theory,
we propose a simple and parallel way to improve SGD estimates for variational i
nference. The approach is theoretically motivated and comes with a diagnostic fo
r convergence and a novel stopping rule, which is robust to noisy objective
functions evaluations. We show empirically, the new workflow works well on a di
verse set of models and datasets, or warns if the stochastic optimization fails
or if the used variational distribution is not good.
**********

Discovering conflicting groups in signed networks
Ruo-Chun Tzeng, Bruno Ordozgoiti, Aristides Gionis
Requests for name changes in the electronic proceedings will be accepted with 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 Some Popular Gaussian Graphical Models without Condition Number Bounds
Jonathan Kelner, Frederic Koehler, Raghu Meka, Ankur Moitra
Gaussian Graphical Models (GGMs) have wide-ranging applications in machine learn
ing and the natural and social sciences. In most of the settings in which they a
re applied, the number of observed samples is much smaller than the dimension an
d they are assumed to be sparse. While there are a variety of algorithms (e.g. G
raphical Lasso, CLIME) that provably recover the graph structure with a logarith
mic number of samples, to do so they require various assumptions on the well-con
ditioning of the precision matrix that are not information-theoretically necessa
ry.
**********

Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserv
ed Confounding
Victor Veitch, Anisha Zaveri
It is a truth universally acknowledged that an observed association without know
n mechanism must be in want of a causal estimate. Causal estimates from observat
ional data will be biased in the presence of ‘unobserved confounding’. However,
we might hope that the influence of unobserved confounders is weak relative to a
‘large’ estimated effect. The purpose of this paper is to develop Austen plots,
a sensitivity analysis tool to aid such judgments by making it easier to reason
about potential bias induced by unobserved confounding. We formalize confoundin
g strength in terms of how strongly the unobserved confounding influences treatm
ent assignment and outcome. For a target level of bias, an Austen plot shows the
minimum values of treatment and outcome influence required to induce that level
of bias. Austen plots generalize the classic sensitivity analysis approach of I
mbens [Imb03]. Critically, Austen plots allow any approach for modeling the obse
rved data. We illustrate the tool by assessing biases for several real causal in
ference problems, using a variety of machine learning approaches for the initial

data analysis. Code, demo data, and a tutorial are available at github.com/anishazaveri/austen_plots.
**********
Mix and Match: An Optimistic Tree-Search Approach for Learning Models from Mixture Distributions
Matthew Faw, Rajat Sen, Karthikeyan Shanmugam, Constantine Caramanis, Sanjay Shakkottai
We consider a covariate shift problem where one has access to several different training datasets for the same learning problem and a small validation set which possibly differs from all the individual training distributions. The distribution shift is due, in part, to \emph{unobserved} features in the datasets. The objective, then, is to find the best mixture distribution over the training datasets (with only observed features) such that training a learning algorithm using this mixture has the best validation performance. Our proposed algorithm, \textsf{Mix\&Match}, combines stochastic gradient descent (SGD) with optimistic tree search and model re-use (evolving partially trained models with samples from different mixture distributions) over the space of mixtures, for this task. We prove a novel high probability bound on the final SGD iterate without relying on a global gradient norm bound, and use it to show the advantages of model re-use. Additionally, we provide simple regret guarantees for our algorithm with respect to recovering the optimal mixture, given a total budget of SGD evaluations. Finally, we validate our algorithm on two real-world datasets.
**********
Understanding Double Descent Requires A Fine-Grained Bias-Variance Decomposition
Ben Adlam, Jeffrey Pennington
Classical learning theory suggests that the optimal generalization performance of a machine learning model should occur at an intermediate model complexity, with simpler models exhibiting high bias and more complex models exhibiting high variance of the predictive function. However, such a simple trade-off does not adequately describe deep learning models that simultaneously attain low bias and variance in the heavily overparameterized regime. A primary obstacle in explaining this behavior is that deep learning algorithms typically involve multiple sources of randomness whose individual contributions are not visible in the total variance. To enable fine-grained analysis, we describe an interpretable, symmetric decomposition of the variance into terms associated with the randomness from sampling, initialization, and the labels. Moreover, we compute the high-dimensional asymptotic behavior of this decomposition for random feature kernel regression, and analyze the strikingly rich phenomenology that arises. We find that the bias decreases monotonically with the network width, but the variance terms exhibit non-monotonic behavior and can diverge at the interpolation boundary, even in the absence of label noise. The divergence is caused by the interaction between sampling and initialization and can therefore be eliminated by marginalizing over samples (i.e. bagging) or over the initial parameters (i.e. ensemble learning).
**********
VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain
Jinsung Yoon, Yao Zhang, James Jordon, Mihaela van der Schaar
Self- and semi-supervised learning frameworks have made significant progress in training machine learning models with limited labeled data in image and language domains. These methods heavily rely on the unique structure in the domain datasets (such as spatial relationships in images or semantic relationships in language). They are not adaptable to general tabular data which does not have the same explicit structure as image and language data. In this paper, we fill this gap by proposing novel self- and semi-supervised learning frameworks for tabular data, which we refer to collectively as VIME (Value Imputation and Mask Estimation). We create a novel pretext task of estimating mask vectors from corrupted tabular data in addition to the reconstruction pretext task for self-supervised learning. We also introduce a novel tabular data augmentation method for self- and semi-supervised learning frameworks. In experiments, we evaluate the proposed fram

ework in multiple tabular datasets from various application domains, such as gen omics and clinical data. VIME exceeds state-of-the-art performance in comparison to the existing baseline methods.
**********

The Smoothed Possibility of Social Choice
Lirong Xia
We develop a framework that leverages the smoothed complexity analysis by Spielm an and Teng to circumvent paradoxes and impossibility theorems in social choice, motivated by modern applications of social choice powered by AI and ML. For Con drocet's paradox, we prove that the smoothed likelihood of the paradox either va nishes at an exponential rate as the number of agents increases, or does not van ish at all. For the ANR impossibility on the non-existence of voting rules that simultaneously satisfy anonymity, neutrality, and resolvability, we characterize the rate for the impossibility to vanish, to be either polynomially fast or exp onentially fast. We also propose a novel easy-to-compute tie-breaking mechanism that optimally preserves anonymity and neutrality for even number of alternative s in natural settings. Our results illustrate the smoothed possibility of social choice—even though the paradox and the impossibility theorem hold in the worst case, they may not be a big concern in practice.
**********

A Decentralized Parallel Algorithm for Training Generative Adversarial Nets
Mingrui Liu, Wei Zhang, Youssef Mroueh, Xiaodong Cui, Jarret Ross, Tianbao Yang, Payel Das
Generative Adversarial Networks (GANs) are a powerful class of generative models in the deep learning community. Current practice on large-scale GAN training ut ilizes large models and distributed large-batch training strategies, and is impl emented on deep learning frameworks (e.g., TensorFlow, PyTorch, etc.) designed i n a centralized manner. In the centralized network topology, every worker needs to either directly communicate with the central node or indirectly communicate w ith all other workers in every iteration. However, when the network bandwidth is low or network latency is high, the performance would be significantly degraded . Despite recent progress on decentralized algorithms for training deep neural n etworks, it remains unclear whether it is possible to train GANs in a decentrali zed manner. The main difficulty lies at handling the nonconvex-nonconcave min-ma x optimization and the decentralized communication simultaneously. In this paper , we address this difficulty by designing the \textbf{first gradient-based decen tralized parallel algorithm} which allows workers to have multiple rounds of com munications in one iteration and to update the discriminator and generator simul taneously, and this design makes it amenable for the convergence analysis of the proposed decentralized algorithm. Theoretically, our proposed decentralized alg orithm is able to solve a class of non-convex non-concave min-max problems with provable non-asymptotic convergence to first-order stationary point. Experimenta l results on GANs demonstrate the effectiveness of the proposed algorithm.
**********

Phase retrieval in high dimensions: Statistical and computational phase transiti ons
Antoine Maillard, Bruno Loureiro, Florent Krzakala, Lenka Zdeborová
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Fair Performance Metric Elicitation
Gaurush Hiranandani, Harikrishna Narasimhan, Sanmi Koyejo
What is a fair performance metric? We consider the choice of fairness metrics th rough the lens of metric elicitation -- a principled framework for selecting per formance metrics that best reflect implicit preferences. The use of metric elici tation enables a practitioner to tune the performance and fairness metrics to th e task, context, and population at hand. Specifically, we propose a novel strat egy to elicit group-fair performance metrics for multiclass classification probl

ems with multiple sensitive groups that also includes selecting the trade-off be
tween predictive performance and fairness violation. The proposed elicitation st
rategy requires only relative preference feedback and is robust to both finite s
ample and feedback noise.
**********
Hybrid Variance-Reduced SGD Algorithms For Minimax Problems with Nonconvex-Linea
r Function
Quoc Tran Dinh, Deyi Liu, Lam Nguyen
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Belief-Dependent Macro-Action Discovery in POMDPs using the Value of Information
Genevieve Flaspohler, Nicholas A. Roy, John W. Fisher III
This work introduces macro-action discovery using value-of-information (VoI) for
 robust and efficient planning in partially observable Markov decision processes
 (POMDPs). POMDPs are a powerful framework for planning under uncertainty. Previ
ous approaches have used high-level macro-actions within POMDP policies to reduc
e planning complexity. However, macro-action design is often heuristic and rarel
y comes with performance guarantees.  Here, we present a method for extracting b
elief-dependent, variable-length macro-actions directly from a low-level POMDP m
odel.  We construct macro-actions by chaining sequences of open-loop actions tog
ether when the task-specific value of information (VoI) --- the change in expect
ed task performance caused by observations in the current planning iteration ---
 is low. Importantly, we provide performance guarantees on the resulting VoI mac
ro-action policies in the form of bounded regret relative to the optimal policy.
 In simulated tracking experiments, we achieve higher reward than both closed-lo
op and hand-coded macro-action baselines, selectively using VoI macro-actions to
 reduce planning complexity while maintaining near-optimal task performance.
**********
Soft Contrastive Learning for Visual Localization
Janine Thoma, Danda Pani Paudel, Luc V. Gool
Localization by image retrieval is inexpensive and scalable due to simple mappin
g and matching techniques. Such localization, however, depends upon the quality
of image features often obtained using Contrastive learning frameworks. Most con
trastive learning strategies opt for features to distinguish different classes.
In the context of localization, however, there is no natural definition of class
es. Therefore, images are usually artificially separated into positive and negat
ive classes, with respect to the chosen anchor images, based on some geometric p
roximity measure. In this paper, we show why such divisions are problematic for
learning localization features. We argue that any artificial division based on s
ome proximity measure is undesirable, due to the inherently ambiguous supervisio
n for images near proximity threshold. To this end, we propose a novel technique
 that uses soft positive/negative assignments of images for contrastive learning
, avoiding the aforementioned problem. Our soft assignment makes a gradual disti
nction between close and far images in both geometric and feature spaces. Experi
ments on four large-scale benchmark datasets demonstrate the superiority of the
proposed soft contrastive learning over the state-of-the-art method for retrieva
l-based visual localization.
**********
Fine-Grained Dynamic Head for Object Detection
Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Hongbin Sun, Jian Sun, Nanning Z
heng
The Feature Pyramid Network (FPN) presents a remarkable approach to alleviate th
e scale variance in object representation by performing instance-level assignmen
ts. Nevertheless, this strategy ignores the distinct characteristics of differen
t sub-regions in an instance. To this end, we propose a fine-grained dynamic hea
d to conditionally select a pixel-level combination of FPN features from differe
nt scales for each instance, which further releases the ability of multi-scale f

eature representation. Moreover, we design a spatial gate with the new activatio n function to reduce computational complexity dramatically through spatially spa rse convolutions. Extensive experiments demonstrate the effectiveness and effici ency of the proposed method on several state-of-the-art detection benchmarks. Co de is available at https://github.com/StevenGrove/DynamicHead.
**********

LoCo: Local Contrastive Representation Learning
Yuwen Xiong, Mengye Ren, Raquel Urtasun
Deep neural nets typically perform end-to-end backpropagation to learn the weigh ts, a procedure that creates synchronization constraints in the weight update st ep across layers and is not biologically plausible. Recent advances in unsupervi sed contrastive representation learning invite the question of whether a learnin g algorithm can also be made local, that is, the updates of lower layers do not directly depend on the computation of upper layers. While Greedy InfoMax separat ely learns each block with a local objective, we found that it consistently hurt s readout accuracy in state-of-the-art unsupervised contrastive learning algorit hms, possibly due to the greedy objective as well as gradient isolation. In this work, we discover that by overlapping local blocks stacking on top of each othe r, we effectively increase the decoder depth and allow upper blocks to implicitl y send feedbacks to lower blocks. This simple design closes the performance gap between local learning and end-to-end contrastive learning algorithms for the fi rst time. Aside from standard ImageNet experiments, we also show results on comp lex downstream tasks such as object detection and instance segmentation directly using readout features.
**********

Modeling and Optimization Trade-off in Meta-learning
Katelyn Gao, Ozan Sener
By searching for shared inductive biases across tasks, meta-learning promises to accelerate learning on novel tasks, but with the cost of solving a complex bile vel optimization problem. We introduce and rigorously define the trade-off betwe en accurate modeling and optimization ease in meta-learning.
At one end, classic meta-learning algorithms account for the structure of meta-l earning but solve a complex optimization problem, while at the other end domain randomized search (otherwise known as joint training) ignores the structure of m eta-learning and solves a single level optimization problem.
Taking MAML as the representative meta-learning algorithm, we theoretically char acterize the trade-off for general non-convex risk functions as well as linear r egression, for which we are able to provide explicit bounds on the errors associ ated with modeling and optimization. We also empirically study this trade-off fo r meta-reinforcement learning benchmarks.
**********

SnapBoost: A Heterogeneous Boosting Machine
Thomas Parnell, Andreea Anghel, Ma■gorzata ■azuka, Nikolas Ioannou, Sebastian Ku rella, Peshal Agarwal, Nikolaos Papandreou, Haralampos Pozidis
Modern gradient boosting software frameworks, such as XGBoost and LightGBM, impl ement Newton descent in a functional space. At each boosting iteration, their go al is to find the base hypothesis, selected from some base hypothesis class, tha t is closest to the Newton descent direction in a Euclidean sense. Typically, th e base hypothesis class is fixed to be all binary decision trees up to a given d epth. In this work, we study a Heterogeneous Newton Boosting Machine (HNBM) in w hich the base hypothesis class may vary across boosting iterations. Specifically , at each boosting iteration, the base hypothesis class is chosen, from a fixed set of subclasses, by sampling from a probability distribution. We derive a glob al linear convergence rate for the HNBM under certain assumptions, and show that it agrees with existing rates for Newton's method when the Newton direction can be perfectly fitted by the base hypothesis at each boosting iteration. We then describe a particular realization of a HNBM, SnapBoost, that, at each boosting i teration, randomly selects between either a decision tree of variable depth or a linear regressor with random Fourier features. We describe how SnapBoost is imp lemented, with a focus on the training complexity. Finally, we present experimen

tal results, using OpenML and Kaggle datasets, that show that SnapBoost is able to achieve better generalization loss than competing boosting frameworks, withou t taking significantly longer to tune.
**********

On Adaptive Distance Estimation
Yeshwanth Cherapanamjeri, 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-auth ors prior to requesting a name change in the electronic proceedings.
**********

Stage-wise Conservative Linear Bandits
Ahmadreza Moradipari, Christos Thrampoulidis, Mahnoosh Alizadeh
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

RELATE: Physically Plausible Multi-Object Scene Synthesis Using Structured Laten t Spaces
Sebastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner , Niloy Mitra, Andrea Vedaldi
We present RELATE, a model that learns to generate physically plausible scenes a nd videos of multiple interacting objects.
Similar to other generative approaches, RELATE is trained end-to-end on raw, unl abeled data.
RELATE combines an object-centric GAN formulation with a model that explicitly a ccounts for correlations between individual objects.
This allows the model to generate realistic scenes and videos from a physically- interpretable parameterization.
Furthermore, we show that modeling the object correlation is necessary to learn to disentangle object positions and identity.
We find that RELATE is also amenable to physically realistic scene editing and t hat it significantly outperforms prior art in object-centric scene generation in both synthetic (CLEVR, ShapeStacks) and real-world data (cars).
In addition, in contrast to state-of-the-art methods in object-centric generativ e modeling, RELATE also extends naturally to dynamic scenes and generates videos of high visual fidelity. Source code, datasets and more results are available a t http://geometry.cs.ucl.ac.uk/projects/2020/relate/.
**********

Metric-Free Individual Fairness in Online Learning
Yahav Bechavod, Christopher Jung, Steven Z. Wu
We study an online learning problem subject to the constraint of individual fair ness, which requires that similar individuals are treated similarly. Unlike prio r work on individual fairness, we do not assume the similarity measure among ind ividuals is known, nor do we assume that such measure takes a certain parametric form. Instead, we leverage the existence of an auditor who detects fairness vi olations without enunciating the quantitative measure. In each round, the audito r examines the learner's decisions and attempts to identify a pair of individual s that are treated unfairly by the learner. We provide a general reduction frame work that reduces online classification in our model to standard online classifi cation, which allows us to leverage existing online learning algorithms to achie ve sub-linear regret and number of fairness violations. Surprisingly, in the sto chastic setting where the data are drawn independently from a distribution, we a re also able to establish PAC-style fairness and accuracy generalization guarant ees (Rothblum and Yona (2018)), despite only having access to a very restricted form of fairness feedback. Our fairness generalization bound qualitatively match es the uniform convergence bound of Rothblum and Yona (2018), while also providi ng a meaningful accuracy generalization guarantee. Our results resolve an open q uestion by Gillen et al. (2018) by showing that online learning under an unknown

individual fairness constraint is possible even without assuming a strong param
etric form of the underlying similarity measure.
**********
GreedyFool: Distortion-Aware Sparse Adversarial Attack
Xiaoyi Dong, Dongdong Chen, Jianmin Bao, Chuan Qin, Lu Yuan, Weiming Zhang, Neng
hai Yu, Dong Chen
Modern deep neural networks(DNNs) are vulnerable to adversarial samples. Sparse
adversarial samples are a special branch of adversarial samples that can fool th
e target model by only perturbing a few pixels. The existence of the sparse adve
rsarial attack points out that DNNs are much more vulnerable than people believe
d, which is also a new aspect for analyzing DNNs. However, current sparse advers
arial attack methods still have some shortcomings on both sparsity and invisibil
ity. In this paper, we propose a novel two-stage distortion-aware greedy-based m
ethod dubbed as ''GreedyFool". Specifically, it first selects the most effective
candidate positions to modify by considering both the gradient(for adversary) a
nd the distortion map(for invisibility), then drops some less important points i
n the reduce stage.
Experiments demonstrate that compared with the start-of-the-art method, we only
need to modify 3 times fewer pixels under the same sparse perturbation setting.
For target attack, the success rate of our method is 9.96% higher than the start
-of-the-art method under the same pixel budget.
**********
VAEM: a Deep Generative Model for Heterogeneous Mixed Type Data
Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, C
heng Zhang
Deep generative models often perform poorly in real-world applications due to th
e heterogeneity of natural data sets. Heterogeneity arises from data containing
different types of features (categorical, ordinal, continuous, etc.) and featur
es of the same type having different marginal distributions. We propose an exten
sion of
variational autoencoders (VAEs) called VAEM to handle such heterogeneous data. V
AEM is a deep generative model that is trained in a two stage manner, such that
the first stage provides a more uniform representation of the data to the second
stage, thereby sidestepping the problems caused by heterogeneous data.
We provide extensions of VAEM to handle partially observed data, and demonstrate
its performance in data generation, missing data prediction and sequential feat
ure selection tasks. Our results show that VAEM broadens the range of real-world
applications where deep generative models can be successfully deployed.
**********
RetroXpert: Decompose Retrosynthesis Prediction Like A Chemist
Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, JINYU YANG, Yang Yu,
Junzhou Huang
Retrosynthesis is the process of recursively decomposing target molecules into a
vailable building blocks. It plays an important role in solving problems in orga
nic synthesis planning. To automate or assist in the retrosynthesis analysis, va
rious retrosynthesis prediction algorithms have been proposed. However, most of
them are cumbersome and lack interpretability about their predictions. In this p
aper, we devise a novel template-free algorithm for automatic retrosynthetic exp
ansion inspired by how chemists approach retrosynthesis prediction. Our method d
isassembles retrosynthesis into two steps: i) identify the potential reaction ce
nter of the target molecule through a novel graph neural network and generate in
termediate synthons, and ii) generate the reactants associated with synthons via
a robust reactant generation model. While outperforming the state-of-the-art ba
selines by a significant margin, our model also provides chemically reasonable i
nterpretation.
**********
Sample-Efficient Optimization in the Latent Space of Deep Generative Models via
Weighted Retraining
Austin Tripp, Erik Daxberger, José Miguel Hernández-Lobato
Many important problems in science and engineering, such as drug design, involve

optimizing an expensive black-box objective function over a complex, high-dimen
sional, and structured input space. Although machine learning techniques have sh
own promise in solving such problems, existing approaches substantially lack sam
ple efficiency. We introduce an improved method for efficient black-box optimiza
tion, which performs the optimization in the low-dimensional, continuous latent
manifold learned by a deep generative model. In contrast to previous approaches,
we actively steer the generative model to maintain a latent manifold that is hi
ghly useful for efficiently optimizing the objective. We achieve this by periodi
cally retraining the generative model on the data points queried along the optim
ization trajectory, as well as weighting those data points according to their ob
jective function value. This weighted retraining can be easily implemented on to
p of existing methods, and is empirically shown to significantly improve their e
fficiency and performance on synthetic and real-world optimization problems.
**********

Improved Sample Complexity for Incremental Autonomous Exploration in MDPs
Jean Tarbouriech, Matteo Pirotta, Michal Valko, 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.
**********

TinyTL: Reduce Memory, Not Parameters for Efficient On-Device Learning
Han Cai, Chuang Gan, Ligeng Zhu, Song Han
Efficient on-device learning requires a small memory footprint at training time
to fit the tight memory constraint. Existing work solves this problem by reducin
g the number of trainable parameters. However, this doesn't directly translate t
o memory saving since the major bottleneck is the activations, not parameters.
In this work, we present Tiny-Transfer-Learning (TinyTL) for memory-efficient on
-device learning. TinyTL freezes the weights while only learns the memory-effici
ent bias modules, thus no need to store the intermediate activations. To maintai
n the adaptation capacity, we introduce a new memory-efficient bias module, the
lite residual module, to refine the feature extractor by learning small residual
feature maps adding only 3.8% memory overhead. Extensive experiments show that
TinyTL significantly saves the memory (up to 6.5x) with little accuracy loss com
pared to fine-tuning the full network. Compared to fine-tuning the last layer, T
inyTL provides significant accuracy improvements (up to 33.8%) with little memor
y overhead. Furthermore, combined with feature extractor adaptation, TinyTL prov
ides 7.5-12.9x memory saving without sacrificing accuracy compared to fine-tunin
g the full Inception-V3. Code is released at https://github.com/mit-han-lab/tiny
ML/tree/master/tinyTL.
**********

RD$^2$: Reward Decomposition with Representation Decomposition
Zichuan Lin, Derek Yang, Li Zhao, Tao Qin, Guangwen Yang, Tie-Yan Liu
Requests for name changes in the electronic proceedings will be accepted with no
questions asked. However name changes may cause bibliographic tracking issues.
Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Self-paced Contrastive Learning with Hybrid Memory for Domain Adaptive Object Re
-ID
Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, hongsheng Li
Domain adaptive object re-ID aims to transfer the learned knowledge from the lab
eled source domain to the unlabeled target domain to tackle the open-class re-id
entification problems. Although state-of-the-art pseudo-label-based methods have
achieved great success, they did not make full use of all valuable information
because of the domain gap and unsatisfying clustering performance. To solve thes
e problems, we propose a novel self-paced contrastive learning framework with hy
brid memory. The hybrid memory dynamically generates source-domain class-level,
target-domain cluster-level and un-clustered instance-level supervisory signals
for learning feature representations. Different from the conventional contrastiv

e learning strategy, the proposed framework jointly distinguishes source-domain classes, and target-domain clusters and un-clustered instances. Most importantly, the proposed self-paced method gradually creates more reliable clusters to refine the hybrid memory and learning targets, and is shown to be the key to our outstanding performance. Our method outperforms state-of-the-arts on multiple domain adaptation tasks of object re-ID and even boosts the performance on the source domain without any extra annotations. Our generalized version on unsupervised object re-ID surpasses state-of-the-art algorithms by considerable 16.7% and 7.9% on Market-1501 and MSMT17 benchmarks.
**********

Fairness constraints can help exact inference in structured prediction
Kevin Bello, 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.
**********

Instance-based Generalization in Reinforcement Learning
Martin Bertran, Natalia Martinez, Mariano Phielipp, Guillermo Sapiro
Agents trained via deep reinforcement learning (RL) routinely fail to generalize to unseen environments, even when these share the same underlying dynamics as the training levels. Understanding the generalization properties of RL is one of the challenges of modern machine learning. Towards this goal, we analyze policy learning in the context of Partially Observable Markov Decision Processes (POMDPs) and formalize the dynamics of training levels as instances. We prove that, independently of the exploration strategy, reusing instances introduces signi■cant changes on the effective Markov dynamics the agent observes during training. Maximizing expected rewards impacts the learned belief state of the agent by inducing undesired instance-speci■c speed-running policies instead of generalizable ones, which are sub-optimal on the training set.
We provide generalization bounds to the value gap in train and test environments based on the number of training instances, and use insights based on these to improve performance on unseen levels. We propose training a shared belief representation over an ensemble of specialized policies, from which we compute a consensus policy that is used for data collection, disallowing instance-speci■c exploitation. We experimentally validate our theory, observations, and the proposed computational solution over the CoinRun benchmark.
**********

Smooth And Consistent Probabilistic Regression Trees
Sami Alkhoury, Emilie Devijver, Marianne Clausel, Myriam Tami, Eric Gaussier, georges Oppenheim
We propose here a generalization of regression trees, referred to as Probabilistic Regression (PR) trees, that adapt to the smoothness of the prediction function relating input and output variables while preserving the interpretability of the prediction and being robust to noise. In PR trees, an observation is associated to all regions of a tree through a probability distribution that reflects how far the observation is to a region. We show that such trees are consistent, meaning that their error tends to 0 when the sample size tends to infinity, a property that has not been established for similar, previous proposals as Soft trees and Smooth Transition Regression trees. We further explain how PR trees can be used in different ensemble methods, namely Random Forests and Gradient Boosted Trees. Lastly, we assess their performance through extensive experiments that illustrate their benefits in terms of performance, interpretability and robustness to noise.
**********

Computing Valid p-value for Optimal Changepoint by Selective Inference using Dynamic Programming
Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, Ichiro Takeuchi
Although there is a vast body of literature related to methods for detecting change-points (CPs), less attention has been paid to assessing the statistical reli

ability of the detected CPs. In this paper, we introduce a novel method to perfo
rm statistical inference on the significance of the CPs, estimated by a Dynamic
Programming (DP)-based optimal CP detection algorithm. Our main idea is to emplo
y a Selective Inference (SI) approach---a new statistical inference framework th
at has recently received a lot of attention---to compute exact (non-asymptotic)
valid p-values for the detected optimal CPs. Although it is well-known that SI h
as low statistical power because of over-conditioning, we address this drawback
by introducing a novel method called parametric DP, which enables SI to be condu
cted with the minimum amount of conditioning, leading to high statistical power.
We conduct experiments on both synthetic and real-world datasets, through which
we offer evidence that our proposed method is more powerful than existing metho
ds, has decent performance in terms of computational efficiency, and provides go
od results in many practical applications.
**********

Factorized Neural Processes for Neural Processes: K-Shot Prediction of Neural Re
sponses
Ronald (James) Cotton, Fabian Sinz, Andreas Tolias
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Winning the Lottery with Continuous Sparsification
Pedro Savarese, Hugo Silva, Michael Maire
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Adversarial robustness via robust low rank representations
Pranjal Awasthi, Himanshu Jain, Ankit Singh Rawat, Aravindan Vijayaraghavan
Adversarial robustness measures the susceptibility of a classifier to impercepti
ble perturbations made to the inputs at test time. In this work we highlight the
benefits of natural low rank representations that often exist for real data suc
h as images, for training neural networks with certified robustness guarantees.
**********

Joints in Random Forests
Alvaro Correia, Robert Peharz, Cassio P. de Campos
Decision Trees (DTs) and Random Forests (RFs) are powerful discriminative learne
rs and tools of central importance to the everyday machine learning practitioner
and data scientist. Due to their discriminative nature, however, they lack prin
cipled methods to process inputs with missing features or to detect outliers, wh
ich requires pairing them with imputation techniques or a separate generative mo
del. In this paper, we demonstrate that DTs and RFs can naturally be interpreted
as generative models, by drawing a connection to Probabilistic Circuits, a prom
inent class of tractable probabilistic models. This reinterpretation equips them
with a full joint distribution over the feature space and leads to Generative D
ecision Trees (GeDTs) and Generative Forests (GeFs), a family of novel hybrid ge
nerative-discriminative models. This family of models retains the overall charac
teristics of DTs and RFs while additionally being able to handle missing feature
s by means of marginalisation. Under certain assumptions, frequently made for Ba
yes consistency results, we show that consistency in GeDTs and GeFs extend to an
y pattern of missing input features, if missing at random. Empirically, we show
that our models often outperform common routines to treat missing data, such as
K-nearest neighbour imputation, and moreover, that our models can naturally dete
ct outliers by monitoring the marginal probability of input features.
**********

Compositional Generalization by Learning Analytical Expressions
Qian Liu, Shengnan An, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Na
nning Zheng, Dongmei Zhang

Compositional generalization is a basic and essential intellective capability of human beings, which allows us to recombine known parts readily. However, existing neural network based models have been proven to be extremely deficient in such a capability. Inspired by work in cognition which argues compositionality can be captured by variable slots with symbolic functions, we present a refreshing view that connects a memory-augmented neural model with analytical expressions, to achieve compositional generalization. Our model consists of two cooperative neural modules, Composer and Solver, fitting well with the cognitive argument while being able to be trained in an end-to-end manner via a hierarchical reinforcement learning algorithm. Experiments on the well-known benchmark SCAN demonstrate that our model seizes a great ability of compositional generalization, solving all challenges addressed by previous works with 100% accuracies.
**********

JAX MD: A Framework for Differentiable Physics
Samuel Schoenholz, Ekin Dogus Cubuk
We introduce JAX MD, a software package for performing differentiable physics simulations with a focus on molecular dynamics. JAX MD includes a number of statistical physics simulation environments as well as interaction potentials and neural networks that can be integrated into these environments without writing any additional code. Since the simulations themselves are differentiable functions, entire trajectories can be differentiated to perform meta-optimization. These features are built on primitive operations, such as spatial partitioning, that allow simulations to scale to hundreds-of-thousands of particles on a single GPU. These primitives are flexible enough that they can be used to scale up workloads outside of molecular dynamics. We present several examples that highlight the features of JAX MD including: integration of graph neural networks into traditional simulations, meta-optimization through minimization of particle packings, and a multi-agent flocking simulation. JAX MD is available at www.github.com/google/jax-md.
**********

An implicit function learning approach for parametric modal regression
Yangchen Pan, Ehsan Imani, Amir-massoud Farahmand, Martha White
For multi-valued functions---such as when the conditional distribution on targets given the inputs is multi-modal---standard regression approaches are not always desirable because they provide the conditional mean. Modal regression algorithms address this issue by instead finding the conditional mode(s). Most, however, are nonparametric approaches and so can be difficult to scale. Further, parametric approximators, like neural networks, facilitate learning complex relationships between inputs and targets. In this work, we propose a parametric modal regression algorithm. We use the implicit function theorem to develop an objective, for learning a joint function over inputs and targets. We empirically demonstrate on several synthetic problems that our method (i) can learn multi-valued functions and produce the conditional modes, (ii) scales well to high-dimensional inputs, and (iii) can even be more effective for certain uni-modal problems, particularly for high-frequency functions. We demonstrate that our method is competitive in a real-world modal regression problem and two regular regression datasets.
**********

SDF-SRN: Learning Signed Distance 3D Object Reconstruction from Static Images
Chen-Hsuan Lin, Chaoyang Wang, Simon Lucey
Dense 3D object reconstruction from a single image has recently witnessed remarkable advances, but supervising neural networks with ground-truth 3D shapes is impractical due to the laborious process of creating paired image-shape datasets. Recent efforts have turned to learning 3D reconstruction without 3D supervision from RGB images with annotated 2D silhouettes, dramatically reducing the cost and effort of annotation. These techniques, however, remain impractical as they still require multi-view annotations of the same object instance during training. As a result, most experimental efforts to date have been limited to synthetic datasets.
In this paper, we address this issue and propose SDF-SRN, an approach that requires only a single view of objects at training time, offering greater utility for

real-world scenarios. SDF-SRN learns implicit 3D shape representations to handle arbitrary shape topologies that may exist in the datasets. To this end, we derive a novel differentiable rendering formulation for learning signed distance functions (SDF) from 2D silhouettes. Our method outperforms the state of the art under challenging single-view supervision settings on both synthetic and real-world datasets.

**********

Coresets for Robust Training of Deep Neural Networks against Noisy Labels

Baharan Mirzasoleiman, Kaidi Cao, Jure Leskovec

Modern neural networks have the capacity to overfit noisy labels frequently found in real-world datasets. Although great progress has been made, existing techniques are very limited in providing theoretical guarantees for the performance of the neural networks trained with noisy labels. To tackle this challenge, we propose a novel approach with strong theoretical guarantees for robust training of neural networks trained with noisy labels. The key idea behind our method is to select subsets of clean data points that provide an approximately low-rank Jacobian matrix. We then prove that gradient descent applied to the subsets cannot overfit the noisy labels, without regularization or early stopping. Our extensive experiments corroborate our theory and demonstrate that deep networks trained on our subsets achieve a significantly superior performance, e.g., 7% increase in accuracy on mini Webvision with 50% noisy labels, compared to state-of-the art.

**********

Adapting to Misspecification in Contextual Bandits

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, Julian Zimmert

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

**********

Convergence of Meta-Learning with Task-Specific Adaptation over Partial Parameters

Kaiyi Ji, Jason D. Lee, Yingbin Liang, H. Vincent Poor

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

**********

MetaPerturb: Transferable Regularizer for Heterogeneous Tasks and Architectures

Jeong Un Ryu, JaeWoong Shin, Hae Beom Lee, Sung Ju Hwang

Regularization and transfer learning are two popular techniques to enhance model generalization on unseen data, which is a fundamental problem of machine learning. Regularization techniques are versatile, as they are task- and architecture-agnostic, but they do not exploit a large amount of data available. Transfer learning methods learn to transfer knowledge from one domain to another, but may not generalize across tasks and architectures, and may introduce new training cost for adapting to the target task. To bridge the gap between the two, we propose a transferable perturbation, MetaPerturb, which is meta-learned to improve generalization performance on unseen data. MetaPerturb is implemented as a set-based lightweight network that is agnostic to the size and the order of the input, which is shared across the layers. Then, we propose a meta-learning framework, to jointly train the perturbation function over heterogeneous tasks in parallel. As MetaPerturb is a set-function trained over diverse distributions across layers and tasks, it can generalize to heterogeneous tasks and architectures. We validate the efficacy and generality of MetaPerturb trained on a specific source domain and architecture, by applying it to the training of diverse neural architectures on heterogeneous target datasets against various regularizers and fine-tuning. The results show that the networks trained with MetaPerturb significantly outperform the baselines on most of the tasks and architectures, with a negligible increase in the parameter size and no hyperparameters to tune.

**********

Learning to solve TV regularised problems with unrolled algorithms
Hamza Cherkaoui, Jeremias Sulam, Thomas Moreau
Total Variation (TV) is a popular regularization strategy that promotes piece-wise constant signals by constraining the ■1-norm of the first order derivative of the estimated signal. The resulting optimization problem is usually solved using iterative algorithms such as proximal gradient descent, primal-dual algorithms or ADMM. However, such methods can require a very large number of iterations to converge to a suitable solution. In this paper, we accelerate such iterative algorithms by unfolding proximal gradient descent solvers in order to learn their parameters for 1D TV regularized problems. While this could be done using the synthesis formulation, we demonstrate that this leads to slower performances. The main difficulty in applying such methods in the analysis formulation lies in proposing a way to compute the derivatives through the proximal operator. As our main contribution, we develop and characterize two approaches to do so, describe their benefits and limitations, and discuss the regime where they can actually improve over iterative procedures. We validate those findings with experiments on synthetic and real data.
**********

Object-Centric Learning with Slot Attention
Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, Thomas Kipf
Learning object-centric representations of complex scenes is a promising step towards enabling efficient abstract reasoning from low-level perceptual features. Yet, most deep learning approaches learn distributed representations that do not capture the compositional properties of natural scenes. In this paper, we present the Slot Attention module, an architectural component that interfaces with perceptual representations such as the output of a convolutional neural network and produces a set of task-dependent abstract representations which we call slots. These slots are exchangeable and can bind to any object in the input by specializing through a competitive procedure over multiple rounds of attention. We empirically demonstrate that Slot Attention can extract object-centric representations that enable generalization to unseen compositions when trained on unsupervised object discovery and supervised property prediction tasks.
**********

Improving robustness against common corruptions by covariate shift adaptation
Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, Matthias Bethge
Today’s state-of-the-art machine vision models are vulnerable to image corruptions like blurring or compression artefacts, limiting their performance in many real-world applications. We here argue that popular benchmarks to measure model robustness against common corruptions (like ImageNet-C) underestimate model robustness in many (but not all) application scenarios. The key insight is that in many scenarios, multiple unlabeled examples of the corruptions are available and can be used for unsupervised online adaptation. Replacing the activation statistics estimated by batch normalization on the training set with the statistics of the corrupted images consistently improves the robustness across 25 different popular computer vision models. Using the corrected statistics, ResNet-50 reaches 62.2% mCE on ImageNet-C compared to 76.7% without adaptation. With the more robust DeepAugment+AugMix model, we improve the state of the art achieved by a ResNet50 model up to date from 53.6% mCE to 45.4% mCE. Even adapting to a single sample improves robustness for the ResNet-50 and AugMix models, and 32 samples are sufficient to improve the current state of the art for a ResNet-50 architecture. We argue that results with adapted statistics should be included whenever reporting scores in corruption benchmarks and other out-of-distribution generalization settings.
**********

Deep Smoothing of the Implied Volatility Surface
Damien Ackerer, Natasa Tagasovska, Thibault Vatter
We present a neural network (NN) approach to fit and predict implied volatility surfaces (IVSs).

Atypically to standard NN applications, financial industry practitioners use suc
h models equally to replicate market prices and to value other financial instrum
ents.
In other words, low training losses are as important as generalization capabilit
ies.
Importantly, IVS models need to generate realistic arbitrage-free option prices,
 meaning that no portfolio can lead to risk-free profits.
We propose an approach guaranteeing the absence of arbitrage opportunities by pe
nalizing the loss using soft constraints.
Furthermore, our method can be combined with standard IVS models in quantitative
 finance, thus providing a NN-based correction when such models fail at replicat
ing observed market prices.
This lets practitioners use our approach as a plug-in on top of classical method
s.
Empirical results show that this approach is particularly useful when only spars
e or erroneous data are available.
We also quantify the uncertainty of the model predictions in regions with few or
 no observations.
We further explore how deeper NNs improve over shallower ones, as well as other
properties of the network architecture.
We benchmark our method against standard IVS models.
By evaluating our method on both training sets, and testing sets, namely, we hig
hlight both their capacity to reproduce observed prices and predict new ones.
**********

Probabilistic Inference with Algebraic Constraints: Theoretical Limits and Pract
ical Approximations

Zhe Zeng, Paolo Morettin, Fanqi Yan, Antonio Vergari, Guy Van den Broeck

Weighted model integration (WMI) is a framework to perform advanced probabilisti
c inference on hybrid domains, i.e., on distributions over mixed continuous-disc
rete random variables and in presence of complex logical and arithmetic constrai
nts. In this work, we advance the WMI framework on both the theoretical and algo
rithmic side. First, we exactly trace the boundaries of tractability for WMI inf
erence by proving that to be amenable to exact and efficient inference a WMI pro
blem has to posses a tree-shaped structure with logarithmic diameter. While this
 result deepens our theoretical understanding of WMI it hinders the practical ap
plicability of exact WMI solvers to real-world problems. To overcome this, we pr
opose the first approximate WMI solver that does not resort to sampling, but per
forms exact inference on one approximate models. Our solution performs message p
assing in a relaxed problem structure iteratively to recover certain lost depend
encies and, as our experiments suggest, is competitive with other SOTA WMI solve
rs.
**********

Provable Online CP/PARAFAC Decomposition of a Structured Tensor via Dictionary L
earning

Sirisha Rambhatla, Xingguo Li, Jarvis Haupt

We consider the problem of factorizing a structured 3-way tensor into its consti
tuent Canonical Polyadic (CP) factors. This decomposition, which can be viewed a
s a generalization of singular value decomposition (SVD) for tensors, reveals ho
w the tensor dimensions (features) interact with each other.  However, since the
 factors are a priori unknown, the corresponding optimization problems are inher
ently non-convex. The existing guaranteed algorithms which handle this non-conve
xity incur an irreducible error (bias), and only apply to cases where all factor
s have the same structure. To this end, we develop a provable algorithm for onli
ne structured tensor factorization, wherein one of the factors obeys some incohe
rence conditions, and the others are sparse. Specifically we show that, under so
me relatively mild conditions on initialization, rank, and sparsity, our algorit
hm recovers the factors exactly (up to scaling and permutation) at a linear rate
. Complementary to our theoretical results, our synthetic and real-world data ev
aluations showcase superior performance compared to related techniques.
**********

Look-ahead Meta Learning for Continual Learning
Gunshi Gupta, Karmesh Yadav, Liam Paull
The continual learning problem involves training models with limited capacity to perform well on a set of an unknown number of sequentially arriving tasks. While meta-learning shows great potential for reducing interference between old and new tasks, the current training procedures tend to be either slow or offline, and sensitive to many hyper-parameters. In this work, we propose Look-ahead MAML (La-MAML), a fast optimisation-based meta-learning algorithm for online-continual learning, aided by a small episodic memory. By incorporating the modulation of per-parameter learning rates in our meta-learning update, our approach also allows us to draw connections to and exploit prior work on hypergradients and meta-descent. This provides a more flexible and efficient way to mitigate catastrophic forgetting compared to conventional prior-based methods.
La-MAML achieves performance superior to other replay-based, prior-based and meta-learning based approaches for continual learning on real-world visual classification benchmarks.
**********

A polynomial-time algorithm for learning nonparametric causal graphs
Ming Gao, Yi Ding, Bryon Aragam
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Sparse Learning with CART
Jason Klusowski
Decision trees with binary splits are popularly constructed using Classification and Regression Trees (CART) methodology. For regression models, this approach recursively divides the data into two near-homogenous daughter nodes according to a split point that maximizes the reduction in sum of squares error (the impurity) along a particular variable. This paper aims to study the statistical properties of regression trees constructed with CART. In doing so, we find that the training error is governed by the Pearson correlation between the optimal decision stump and response data in each node, which we bound by constructing a prior distribution on the split points and solving a nonlinear optimization problem. We leverage this connection between the training error and Pearson correlation to show that CART with cost-complexity pruning achieves an optimal complexity/goodness-of-fit tradeoff when the depth scales with the logarithm of the sample size. Data dependent quantities, which adapt to the dimensionality and latent structure of the regression model, are seen to govern the rates of convergence of the prediction error.
**********

Proximal Mapping for Deep Regularization
Mao Li, Yingyi Ma, Xinhua Zhang
Underpinning the success of deep learning is effective regularizations that allow a variety of priors in data to be modeled. For example, robustness to adversarial perturbations, and correlations between multiple modalities. However, most regularizers are specified in terms of hidden layer outputs, which are not themselves optimization variables. In contrast to prevalent methods that optimize them indirectly through model weights, we propose inserting proximal mapping as a new layer to the deep network, which directly and explicitly produces well regularized hidden layer outputs. The resulting technique is shown well connected to kernel warping and dropout, and novel algorithms were developed for robust temporal learning and multiview modeling, both outperforming state-of-the-art methods.
**********

Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models
Andrew Jesson, Sören Mindermann, Uri Shalit, Yarin Gal
Recommending the best course of action for an individual is a major application of individual-level causal effect estimation. This application is often needed in safety-critical domains such as healthcare, where estimating and communicating

uncertainty to decision-makers is crucial. We introduce a practical approach fo
r integrating uncertainty estimation into a class of state-of-the-art neural net
work methods used for individual-level causal estimates. We show that our method
s enable us to deal gracefully with situations of "no-overlap", common in high-d
imensional data, where standard applications of causal effect approaches fail. F
urther, our methods allow us to handle covariate shift, where the train and test
distributions differ, common when systems are deployed in practice. We show tha
t when such a covariate shift occurs, correctly modeling uncertainty can keep us
from giving overconfident and potentially harmful recommendations. We demonstra
te our methodology with a range of state-of-the-art models. Under both covariate
shift and lack of overlap, our uncertainty-equipped methods can alert decision
makers when predictions are not to be trusted while outperforming standard metho
ds that use the propensity score to identify lack of overlap.
**********
Hierarchical Granularity Transfer Learning
Shaobo Min, Hongtao Xie, Hantao Yao, Xuran Deng, Zheng-Jun Zha, Yongdong Zhang
In the real world, object categories usually have a hierarchical granularity tre
e.
Nowadays, most researchers focus on recognizing categories in a specific granula
rity, \emph{e.g.,} basic-level or sub(ordinate)-level. Compared with basic-level
categories, the sub-level categories provide more valuable information, but its
training annotations are harder to acquire. Therefore, an attractive problem is
how to transfer the knowledge learned from basic-level annotations to sub-level
recognition. In this paper, we introduce a new task, named Hierarchical Granula
rity Transfer Learning (HGTL), to recognize sub-level categories with basic-leve
l annotations and semantic descriptions for hierarchical categories. Different f
rom other recognition tasks, HGTL has a serious granularity gap,~\emph{i.e.,} th
e two granularities share an image space but have different category domains, wh
ich impede the knowledge transfer. To this end, we propose a novel Bi-granularit
y Semantic Preserving Network (BigSPN) to bridge the granularity gap for robust
knowledge transfer. Explicitly, BigSPN constructs specific visual encoders for d
ifferent granularities, which are aligned with a shared semantic interpreter via
a novel subordinate entropy loss.  Experiments on three benchmarks with hierarc
hical granularities show that BigSPN is an effective framework for Hierarchical
Granularity Transfer Learning.
**********
Deep active inference agents using Monte-Carlo methods
Zafeirios Fountas, Noor Sajid, Pedro Mediano, Karl Friston
Active inference is a Bayesian framework for understanding biological intelligen
ce. The underlying theory brings together perception and action under one single
imperative: minimizing free energy. However, despite its theoretical utility in
explaining intelligence, computational implementations have been restricted to
low-dimensional and idealized situations. In this paper, we present a neural arc
hitecture for building deep active inference agents operating in complex, contin
uous state-spaces using multiple forms of Monte-Carlo (MC) sampling. For this, w
e introduce a number of techniques, novel to active inference. These include: i)
selecting free-energy-optimal policies via MC tree search, ii) approximating th
is optimal policy distribution via a feed-forward `habitual' network, iii) predi
cting future parameter belief updates using MC dropouts and, finally, iv) optimi
zing state transition precision (a high-end form of attention). Our approach ena
bles agents to learn environmental dynamics efficiently, while maintaining task
performance, in relation to reward-based counterparts. We illustrate this in a n
ew toy environment, based on the dSprites data-set, and demonstrate that active
inference agents automatically create disentangled representations that are apt
for modeling state transitions. In a more complex Animal-AI environment, our age
nts (using the same neural architecture) are able to simulate future state trans
itions and actions (i.e., plan), to evince reward-directed navigation - despite
temporary suspension of visual input. These results show that deep active infere
nce - equipped with MC methods - provides a flexible framework to develop biolog
ically-inspired intelligent agents, with applications in both machine learning a

nd cognitive science.
**********
Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations
Alexander Ritchie, Robert A. Vandermeulen, Clayton Scott
Recent research has established sufficient conditions for finite mixture models to be identifiable from grouped observations. These conditions allow the mixture components to be nonparametric and have substantial (or even total) overlap. This work proposes an algorithm that consistently estimates any identifiable mixture model from grouped observations. Our analysis leverages an oracle inequality for weighted kernel density estimators of the distribution on groups, together with a general result showing that consistent estimation of the distribution on groups implies consistent estimation of mixture components. A practical implementation is provided for paired observations, and the approach is shown to outperform existing methods, especially when mixture components overlap significantly.
**********
Manifold structure in graph embeddings
Patrick Rubin-Delanchy
Statistical analysis of a graph often starts with embedding, the process of representing its nodes as points in space. How to choose the embedding dimension is a nuanced decision in practice, but in theory a notion of true dimension is often available. In spectral embedding, this dimension may be very high. However, this paper shows that existing random graph models, including graphon and other latent position models, predict the data should live near a much lower-dimensional set. One may therefore circumvent the curse of dimensionality by employing methods which exploit hidden manifold structure.
**********
Adaptive Learned Bloom Filter (Ada-BF): Efficient Utilization of the Classifier with Application to Real-Time Information Filtering on the Web
Zhenwei Dai, Anshumali Shrivastava
Recent work suggests improving the performance of Bloom filter by incorporating a machine learning model as a binary classifier. However, such learned Bloom filter does not take full advantage of the predicted probability scores. We propose new algorithms that generalize the learned Bloom filter by using the complete spectrum of the score regions. We prove our algorithms have lower false positive rate (FPR) and memory usage compared with the existing approaches to learned Bloom filter. We also demonstrate the improved performance of our algorithms on real-world information filtering tasks over the web.
**********
MCUNet: Tiny Deep Learning on IoT Devices
Ji Lin, Wei-Ming Chen, Yujun Lin, john cohn, Chuang Gan, Song Han
Machine learning on tiny IoT devices based on microcontroller units (MCU) is appealing but challenging: the memory of microcontrollers is 2-3 orders of magnitude smaller even than mobile phones. We propose MCUNet, a framework that jointly designs the efficient neural architecture (TinyNAS) and the lightweight inference engine (TinyEngine), enabling ImageNet-scale inference on microcontrollers. TinyNAS adopts a two-stage neural architecture search approach that first optimizes the search space to fit the resource constraints, then specializes the network architecture in the optimized search space. TinyNAS can automatically handle diverse constraints (i.e. device, latency, energy, memory) under low search costs. TinyNAS is co-designed with TinyEngine, a memory-efficient inference library to expand the search space and fit a larger model. TinyEngine adapts the memory scheduling according to the overall network topology rather than layer-wise optimization, reducing the memory usage by 3.4×, and accelerating the inference by 1.7-3.3× compared to TF-Lite Micro [3] and CMSIS-NN [28]. MCUNet is the first to achieves >70% ImageNet top1 accuracy on an off-the-shelf commercial microcontroller, using 3.5× less SRAM and 5.7× less Flash compared to quantized MobileNetV2 and ResNet-18. On visual&audio wake words tasks, MCUNet achieves state-of-the-art accuracy and runs 2.4-3.4× faster than Mo- bileNetV2 and ProxylessNAS-based solutions with 3.7-4.1× smaller peak SRAM. Our study suggests that the era of always-

on tiny machine learning on IoT devices has arrived.
**********
In search of robust measures of generalization
Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, Daniel M. Roy
One of the principal scientific challenges in deep learning is explaining generalization, i.e., why the particular way the community now trains networks to achieve small training error also leads to small error on held-out data from the same population. It is widely appreciated that some worst-case theories -- such as those based on the VC dimension of the class of predictors induced by modern neural network architectures -- are unable to explain empirical performance. A large volume of work aims to close this gap, primarily by developing bounds on generalization error, optimization error, and excess risk. When evaluated empirically, however, most of these bounds are numerically vacuous. Focusing on generalization bounds, this work addresses the question of how to evaluate such bounds empirically. Jiang et al. (2020) recently described a large-scale empirical study aimed at uncovering potential causal relationships between bounds/measures and generalization. Building on their study, we highlight where their proposed methods can obscure failures and successes of generalization measures in explaining generalization. We argue that generalization measures should instead be evaluated within the framework of distributional robustness.
**********
Task-agnostic Exploration in Reinforcement Learning
Xuezhou Zhang, Yuzhe Ma, 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.
**********
Multi-task Additive Models for Robust Estimation and Automatic Structure Discovery
Yingjie Wang, Hong Chen, Feng Zheng, Chen Xu, Tieliang Gong, Yanhong Chen
Additive models have attracted much attention for high-dimensional regression estimation and variable selection. However, the existing models are usually limited to the single-task learning framework under the mean squared error (MSE) criterion, where the utilization of variable structure depends heavily on priori knowledge among variables. For high-dimensional observations in real environment, e.g., Coronal Mass Ejections (CMEs) data, the learning performance of previous methods may be degraded seriously due to the complex non-Gaussian noise and the insufficiency of prior knowledge on variable structure. To tackle this problem, we propose a new class of additive models, called Multi-task Additive Models (MAM), by integrating the mode-induced metric, the structure-based regularizer, and additive hypothesis spaces into a bilevel optimization framework. Our approach does not require any priori knowledge of variable structure and suits for high-dimensional data with complex noise, e.g., skewed noise, heavy-tailed noise, and outliers. A smooth iterative optimization algorithm with convergence guarantees is provided to implement MAM efficiently. Experiments on simulations and the CMEs analysis demonstrate the competitive performance of our approach for robust estimation and automatic structure discovery.
**********
Provably Efficient Reward-Agnostic Navigation with Linear Value Iteration
Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, Emma Brunskill
There has been growing progress on theoretical analyses for provably efficient learning in MDPs with linear function approximation, but much of the existing work has made strong assumptions to enable exploration by conventional exploration frameworks. Typically these assumptions are stronger than what is needed to find good solutions in the batch setting. In this work, we show how under a more standard notion of low inherent Bellman error, typically employed in least-square value iteration-style algorithms, we can provide strong PAC guarantees on learning a near optimal value function provided that the linear space is sufficiently `

`explorable''.
We present a computationally tractable algorithm for the reward-free setting and show how it can be used to learn a near optimal policy for any (linear) reward function, which is revealed only once learning has completed. If this reward function is also estimated from the samples gathered during pure exploration, our results also provide same-order PAC guarantees on the performance of the resulting policy for this setting.
**********
Softmax Deep Double Deterministic Policy Gradients
Ling Pan, Qingpeng Cai, Longbo Huang
A widely-used actor-critic reinforcement learning algorithm for continuous control, Deep Deterministic Policy Gradients (DDPG), suffers from the overestimation problem, which can negatively affect the performance. Although the state-of-the-art Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm mitigates the overestimation issue, it can lead to a large underestimation bias. In this paper, we propose to use the Boltzmann softmax operator for value function estimation in continuous control. We first theoretically analyze the softmax operator in continuous action space. Then, we uncover an important property of the softmax operator in actor-critic algorithms, i.e., it helps to smooth the optimization landscape, which sheds new light on the benefits of the operator. We also design two new algorithms, Softmax Deep Deterministic Policy Gradients (SD2) and Softmax Deep Double Deterministic Policy Gradients (SD3), by building the softmax operator upon single and double estimators, which can effectively improve the overestimation and underestimation bias. We conduct extensive experiments on challenging continuous control tasks, and results show that SD3 outperforms state-of-the-art methods.
**********
Online Decision Based Visual Tracking via Reinforcement Learning
ke Song, Wei Zhang, Ran Song, Yibin Li
A deep visual tracker is typically based on either object detection or template matching while each of them is only suitable for a particular group of scenes. It is straightforward to consider fusing them together to pursue more reliable tracking. However, this is not wise as they follow different tracking principles. Unlike previous fusion-based methods, we propose a novel ensemble framework, named DTNet, with an online decision mechanism for visual tracking based on hierarchical reinforcement learning. The decision mechanism substantiates an intelligent switching strategy where the detection and the template trackers have to compete with each other to conduct tracking within different scenes that they are adept in. Besides, we present a novel detection tracker which avoids the common issue of incorrect proposal. Extensive results show that our DTNet achieves state-of-the-art tracking performance as well as good balance between accuracy and efficiency. The project website is available at https://vsislab.github.io/DTNet/.
**********
Efficient Marginalization of Discrete and Structured Latent Variables via Sparsity
Gonçalo Correia, Vlad Niculae, Wilker Aziz, André Martins
Training neural network models with discrete (categorical or structured) latent variables can be computationally challenging, due to the need for marginalization over large or combinatorial sets. To circumvent this issue, one typically resorts to sampling-based approximations of the true marginal, requiring noisy gradient estimators (e.g., score function estimator) or continuous relaxations with lower-variance reparameterized gradients (e.g., Gumbel-Softmax). In this paper, we propose a new training strategy which replaces these estimators by an exact yet efficient marginalization. To achieve this, we parameterize discrete distributions over latent assignments using differentiable sparse mappings: sparsemax and its structured counterparts. In effect, the support of these distributions is greatly reduced, which enables efficient marginalization. We report successful results in three tasks covering a range of latent variable modeling applications: a semisupervised deep generative model, a latent communication game, and a generative model with a bit-vector latent representation. In all cases, we obtain goo

d performance while still achieving the practicality of sampling-based approxima
tions.
**********

DeepI2I: Enabling Deep Hierarchical Image-to-Image Translation by Transferring f
rom GANs
yaxing wang, Lu Yu, Joost van de Weijer
Image-to-image translation has recently achieved remarkable results. But despite
 current success, it suffers from inferior performance when translations between
 classes require large shape changes. We attribute this to the high-resolution b
ottlenecks which are used by current state-of-the-art image-to-image methods.
Therefore, in this work, we propose a novel deep hierarchical Image-to-Image Tra
nslation method, called DeepI2I. We learn a model by leveraging  hierarchical fe
atures: (a) structural information contained in the bottom layers and (b) semant
ic information extracted from the top layers.  To enable the training of deep I2
I models on small datasets, we propose a novel transfer learning method, that tr
ansfers knowledge from pre-trained GANs.  Specifically, we leverage the discrimi
nator of a pre-trained GANs (i.e. BigGAN or StyleGAN) to initialize both the enc
oder and the discriminator and the pre-trained generator to initialize the gener
ator of our model.  Applying knowledge transfer leads to an alignment problem be
tween the encoder and generator. We introduce an adaptor network to address this
.   On many-class image-to-image translation on three datasets (Animal faces, Bi
rds, and Foods) we decrease mFID by at least 35% when compared to the state-of-t
he-art. Furthermore, we qualitatively and quantitatively demonstrate that transf
er learning significantly improves the performance of I2I systems, especially fo
r small datasets.
Finally, we are the first to perform I2I translations for domains with over 100
classes.
**********

Distributional Robustness with IPMs and links to Regularization and GANs
Hisham Husain
Requests for name changes in the electronic proceedings will be accepted with 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 shooting formulation of deep learning
François-Xavier Vialard, Roland Kwitt, Susan Wei, Marc Niethammer
A residual network may be regarded as a discretization of an ordinary differenti
al equation (ODE) which, in the limit of time discretization, defines a continuo
us-depth network. Although important steps have been taken to realize the advant
ages of such continuous formulations, most current techniques assume identical l
ayers. Indeed, existing works throw into relief the myriad difficulties of learn
ing an infinite-dimensional parameter in a continuous-depth neural network. To t
his end, we introduce a shooting formulation which shifts the perspective from p
arameterizing a network layer-by-layer to parameterizing over optimal networks d
escribed only by a set of initial conditions. For scalability, we propose a nove
l particle-ensemble parameterization which fully specifies the optimal weight tr
ajectory of the continuous-depth neural network. Our experiments show that our p
article-ensemble shooting formulation can achieve competitive performance. Final
ly, though the current work is inspired by continuous-depth neural networks, the
 particle-ensemble shooting formulation also applies to discrete-time networks a
nd may lead to a new fertile area of research in deep learning parameterization.
**********

CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Inst
ances
Jihoon Tack, Sangwoo Mo, Jongheon Jeong, Jinwoo Shin
Novelty detection, i.e., identifying whether a given sample is drawn from outsid
e the training distribution, is essential for reliable machine learning. To this
 end, there have been many attempts at learning a representation well-suited for
 novelty detection and designing a score based on such representation. In this p

aper, we propose a simple, yet effective method named contrasting shifted instan
ces (CSI), inspired by the recent success on contrastive learning of visual repr
esentations. Specifically, in addition to contrasting a given sample with other
instances as in conventional contrastive learning methods, our training scheme c
ontrasts the sample with distributionally-shifted augmentations of itself. Based
 on this, we propose a new detection score that is specific to the proposed trai
ning scheme. Our experiments demonstrate the superiority of our method under var
ious novelty detection scenarios, including unlabeled one-class, unlabeled multi
-class and labeled multi-class settings, with various image benchmark datasets.
Code and pre-trained models are available at https://github.com/alinlab/CSI.
**********

Learning Implicit Credit Assignment for Cooperative Multi-Agent Reinforcement Le
arning
Meng Zhou, Ziyu Liu, Pengwei Sui, Yixuan Li, Yuk Ying Chung
We present a multi-agent actor-critic method that aims to implicitly address the
 credit assignment problem under fully cooperative settings. Our key motivation
is that credit assignment among agents may not require an explicit formulation a
s long as (1) the policy gradients derived from a centralized critic carry suffi
cient information for the decentralized agents to maximize their joint action va
lue through optimal cooperation and (2) a sustained level of exploration is enfo
rced throughout training. Under the centralized training with decentralized exec
ution (CTDE) paradigm, we achieve the former by formulating the centralized crit
ic as a hypernetwork such that a latent state representation is integrated into
the policy gradients through its multiplicative association with the stochastic
policies; to achieve the latter, we derive a simple technique called adaptive en
tropy regularization where magnitudes of the entropy gradients are dynamically r
escaled based on the current policy stochasticity to encourage consistent levels
 of exploration. Our algorithm, referred to as LICA, is evaluated on several ben
chmarks including the multi-agent particle environments and a set of challenging
 StarCraft II micromanagement tasks, and we show that LICA significantly outperf
orms previous methods.
**********

MATE: Plugging in Model Awareness to Task Embedding for Meta Learning
Xiaohan Chen, Zhangyang Wang, Siyu Tang, Krikamol Muandet
Meta-learning improves generalization of machine learning models when faced with
 previously unseen tasks by leveraging experiences from different, yet related p
rior tasks. To allow for better generalization, we propose a novel task represen
tation called model-aware task embedding (MATE) that incorporates not only the d
ata distributions of different tasks, but also the complexity of the tasks throu
gh the models used. The task complexity is taken into account by a novel variant
 of kernel mean embedding, combined with an instance-adaptive attention mechanis
m inspired by an SVM-based feature selection algorithm. Together with conditioni
ng layers in deep neural networks, MATE can be easily incorporated into existing
 meta learners as a plug-and-play module. While MATE is widely applicable to gen
eral tasks where the concept of task/environment is involved, we demonstrate its
 effectiveness in few-shot learning by improving a state-of-the-art model consis
tently on two benchmarks. Source codes for this paper are available at https://g
ithub.com/VITA-Group/MATE.
**********

Restless-UCB, an Efficient and Low-complexity Algorithm for Online Restless Band
its
Siwei Wang, Longbo Huang, John C. S. Lui
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Predictive Information Accelerates Learning in RL
Kuang-Huei Lee, Ian Fischer, Anthony Liu, Yijie Guo, Honglak Lee, John Canny, Se
rgio Guadarrama

The Predictive Information is the mutual information between the past and the fu
ture, I(Xpast; Xfuture). We hypothesize that capturing the predictive informatio
n is useful in RL, since the ability to model what will happen next is necessary
 for success on many tasks. To test our hypothesis, we train Soft Actor-Critic (
SAC) agents from pixels with an auxiliary task that learns a compressed represen
tation of the predictive information of the RL environment dynamics using a cont
rastive version of the Conditional Entropy Bottleneck (CEB) objective. We refer
to these as Predictive Information SAC (PI-SAC) agents. We show that PI-SAC agen
ts can substantially improve sample efficiency over challenging baselines on tas
ks from the DM Control suite of continuous control environments. We evaluate PI-
SAC agents by comparing against uncompressed PI-SAC agents, other compressed and
 uncompressed agents, and SAC agents directly trained from pixels. Our implement
ation is given on GitHub.
**********

Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization
Sam Hopkins, Jerry Li, Fred Zhang
We study the problem of estimating the mean of a distribution in high dimensions
 when either the samples are  adversarially corrupted or the distribution is hea
vy-tailed. Recent developments in robust statistics have established  efficient
 and (near) optimal procedures for both settings.  However, the algorithms devel
oped on each side tend to be sophisticated and do not directly transfer to the o
ther, with many of them having ad-hoc or complicated analyses.
**********

High-Fidelity Generative Image Compression
Fabian Mentzer, George D. Toderici, Michael Tschannen, Eirikur Agustsson
We extensively study how to combine Generative Adversarial Networks and learned
compression to obtain a state-of-the-art generative lossy compression system. In
 particular, we investigate normalization layers, generator and discriminator ar
chitectures, training strategies, as well as perceptual losses. In contrast to p
revious work, i) we obtain visually pleasing reconstructions that are perceptual
ly similar to the input, ii) we operate in a broad range of bitrates, and iii) o
ur approach can be applied to high-resolution images. We bridge the gap between
rate-distortion-perception theory and practice by evaluating our approach both q
uantitatively with various perceptual metrics, and with a user study. The study
shows that our method is preferred to previous approaches even if they use more
than 2x the bitrate.
**********

A Statistical Mechanics Framework for Task-Agnostic Sample Design in Machine Lea
rning
Bhavya Kailkhura, Jayaraman Thiagarajan, Qunwei Li, Jize Zhang, Yi Zhou, Timo Br
emer
In this paper, we present a statistical mechanics framework to understand the ef
fect of sampling properties of training data on the generalization gap of machin
e learning (ML) algorithms. We connect the generalization gap to the spatial pro
perties of a sample design characterized by the pair correlation function (PCF).
 In particular, we express generalization gap in terms of the power spectra of t
he sample design and that of the function to be learned. Using this framework, w
e show that space-filling sample designs, such as blue noise and Poisson disk sa
mpling, which optimize spectral properties, outperform random designs in terms o
f the generalization gap and characterize this gain in a closed-form. Our analys
is also sheds light on design principles for constructing optimal task-agnostic
sample designs that minimize the generalization gap. We corroborate our findings
 using regression experiments with neural networks on: a) synthetic functions, a
nd b) a complex scientific simulator for inertial confinement fusion (ICF).
**********

Counterexample-Guided Learning of Monotonic Neural Networks
Aishwarya Sivaraman, Golnoosh Farnadi, Todd Millstein, Guy Van den Broeck
The widespread adoption of deep learning is often attributed to its automatic fe
ature construction with minimal inductive bias. However, in many real-world task
s, the learned function is intended to satisfy domain-specific constraints. We f

ocus on monotonicity constraints, which are common and require that the function 's output increases with increasing values of specific input features. We develo p a counterexample-guided technique to provably enforce monotonicity constraints at prediction time. Additionally, we propose a technique to use monotonicity as an inductive bias for deep learning. It works by iteratively incorporating mono tonicity counterexamples in the learning process. Contrary to prior work in mono tonic learning, we target general ReLU neural networks and do not further restri ct the hypothesis space. We have implemented these techniques in a tool called C OMET. Experiments on real-world datasets demonstrate that our approach achieves state-of-the-art results compared to existing monotonic learners, and can improv e the model quality compared to those that were trained without taking monotonic ity constraints into account.
**********

A Novel Approach for Constrained Optimization in Graphical Models
Sara Rouhani, Tahrima Rahman, Vibhav Gogate
Requests for name changes in the electronic proceedings will be accepted with 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 Deep Networks with One Wide Layer Followed by Pyramidal To pology
Quynh N. Nguyen, Marco Mondelli
Recent works have shown that gradient descent can find a global minimum for over -parameterized neural networks where the widths of all the hidden layers scale p olynomially with N (N being the number of training samples). In this paper, we p rove that, for deep networks, a single layer of width N following the input laye r suffices to ensure a similar guarantee. In particular, all the remaining layer s are allowed to have constant widths, and form a pyramidal topology. We show an application of our result to the widely used Xavier's initialization and obtain an over-parameterization requirement for the single wide layer of order N^2.
**********

On the Trade-off between Adversarial and Backdoor Robustness
Cheng-Hsin Weng, Yan-Ting Lee, Shan-Hung (Brandon) Wu
Deep neural networks are shown to be susceptible to both adversarial attacks and backdoor attacks. Although many defenses against an individual type of the abov e attacks have been proposed, the interactions between the vulnerabilities of a network to both types of attacks have not been carefully investigated yet. In th is paper, we conduct experiments to study whether adversarial robustness and bac kdoor robustness can affect each other and find a trade-off—by increasing the ro bustness of a network to adversarial examples, the network becomes more vulnerab le to backdoor attacks. We then investigate the cause and show how such a trade- off can be exploited for either good or bad purposes. Our findings suggest that future research on defense should take both adversarial and backdoor attacks int o account when designing algorithms or robustness measures to avoid pitfalls and a false sense of security.
**********

Implicit Graph Neural Networks
Fangda Gu, Heng Chang, Wenwu Zhu, Somayeh Sojoudi, Laurent El Ghaoui
Graph Neural Networks (GNNs) are widely used deep learning models that learn mea ningful representations from graph-structured data. Due to the finite nature of the underlying recurrent structure, current GNN methods may struggle to capture long-range dependencies in underlying graphs. To overcome this difficulty, we pr opose a graph learning framework, called Implicit Graph Neural Networks (IGNN), where predictions are based on the solution of a fixed-point equilibrium equatio n involving implicitly defined "state" vectors. We use the Perron-Frobenius theo ry to derive sufficient conditions that ensure well-posedness of the framework. Leveraging implicit differentiation, we derive a tractable projected gradient de scent method to train the framework. Experiments on a comprehensive range of tas ks show that IGNNs consistently capture long-range dependencies and outperform s

tate-of-the-art GNN models.
**********

Rethinking Importance Weighting for Deep Learning under Distribution Shift
Tongtong Fang, Nan Lu, Gang Niu, Masashi Sugiyama
Under distribution shift (DS) where the training data distribution differs from the test one, a powerful technique is importance weighting (IW) which handles DS in two separate steps: weight estimation (WE) estimates the test-over-training density ratio and weighted classification (WC) trains the classifier from weighted training data. However, IW cannot work well on complex data, since WE is incompatible with deep learning. In this paper, we rethink IW and theoretically show it suffers from a circular dependency: we need not only WE for WC, but also WC for WE where a trained deep classifier is used as the feature extractor (FE). To cut off the dependency, we try to pretrain FE from unweighted training data, which leads to biased FE. To overcome the bias, we propose an end-to-end solution dynamic IW that iterates between WE and WC and combines them in a seamless manner, and hence our WE can also enjoy deep networks and stochastic optimizers indirectly. Experiments with two representative types of DS on three popular datasets show that our dynamic IW compares favorably with state-of-the-art methods.
**********

Guiding Deep Molecular Optimization with Genetic Exploration
Sungsoo Ahn, Junsu Kim, Hankook Lee, Jinwoo Shin
De novo molecular design attempts to search over the chemical space for molecules with the desired property. Recently, deep learning has gained considerable attention as a promising approach to solve the problem. In this paper, we propose genetic expert-guided learning (GEGL), a simple yet novel framework for training a deep neural network (DNN) to generate highly-rewarding molecules. Our main idea is to design a "genetic expert improvement" procedure, which generates high-quality targets for imitation learning of the DNN. Extensive experiments show that GEGL significantly improves over state-of-the-art methods. For example, GEGL manages to solve the penalized octanol-water partition coefficient optimization with a score of 31.40, while the best-known score in the literature is 27.22. Besides, for the GuacaMol benchmark with 20 tasks, our method achieves the highest score for 19 tasks, in comparison with state-of-the-art methods, and newly obtains the perfect score for three tasks. Our training code is available at https://github.com/sungsoo-ahn/genetic-expert-guided-learning.
**********

Temporal Spike Sequence Learning via Backpropagation for Deep Spiking Neural Networks
Wenrui Zhang, Peng Li
Spiking neural networks (SNNs) are well suited for spatio-temporal learning and implementations on energy-efficient event-driven neuromorphic processors. However, existing SNN error backpropagation (BP) methods lack proper handling of spiking discontinuities and suffer from low performance compared with the BP methods for traditional artificial neural networks. In addition, a large number of time steps are typically required to achieve decent performance, leading to high latency and rendering spike-based computation unscalable to deep architectures. We present a novel Temporal Spike Sequence Learning Backpropagation (TSSL-BP) method for training deep SNNs, which breaks down error backpropagation across two types of inter-neuron and intra-neuron dependencies and leads to improved temporal learning precision. It captures inter-neuron dependencies through presynaptic firing times by considering the all-or-none characteristics of firing activities and captures intra-neuron dependencies by handling the internal evolution of each neuronal state in time. TSSL-BP efficiently trains deep SNNs within a much shortened temporal window of a few steps while improving the accuracy for various image classification datasets including CIFAR10.
**********

TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation
DONGXU LI, Chenchen Xu, Xin Yu, Kaihao Zhang, Benjamin Swift, Hanna Suominen, Hongdong Li

Sign language translation (SLT) aims to interpret sign video sequences into text-based natural language sentences. Sign videos consist of continuous sequences of sign gestures with no clear boundaries in between. Existing SLT models usually represent sign visual features in a frame-wise manner so as to avoid needing to explicitly segmenting the videos into isolated signs. However, these methods neglect the temporal information of signs and lead to substantial ambiguity in translation. In this paper, we explore the temporal semantic structures of sign videos to learn more discriminative features. To this end, we first present a novel sign video segment representation which takes into account multiple temporal granularities, thus alleviating the need for accurate video segmentation. Taking advantage of the proposed segment representation, we develop a novel hierarchical sign video feature learning method via a temporal semantic pyramid network, called TSPNet. Specifically, TSPNet introduces an inter-scale attention to evaluate and enhance local semantic consistency of sign segments and an intra-scale attention to resolve semantic ambiguity by using non-local video context. Experiments show that our TSPNet outperforms the state-of-the-art with significant improvements on the BLEU score (from 9.58 to 13.41) and ROUGE score (from 31.80 to 34.96) on the largest commonly used SLT dataset. Our implementation is available at https://github.com/verashira/TSPNet.

**********

Neural Topographic Factor Analysis for fMRI Data

Eli Sennesh, Zulqarnain Khan, Yiyu Wang, J Benjamin Hutchinson, Ajay Satpute, Jennifer Dy, Jan-Willem van de Meent

Neuroimaging studies produce gigabytes of spatio-temporal data for a small number of participants and stimuli. Recent work increasingly suggests that the common practice of averaging across participants and stimuli leaves out systematic and meaningful information. We propose Neural Topographic Factor Analysis (NTFA), a probabilistic factor analysis model that infers embeddings for participants and stimuli. These embeddings allow us to reason about differences between participants and stimuli as signal rather than noise. We evaluate NTFA on data from an in-house pilot experiment, as well as two publicly available datasets. We demonstrate that inferring representations for participants and stimuli improves predictive generalization to unseen data when compared to previous topographic methods. We also demonstrate that the inferred latent factor representations are useful for downstream tasks such as multivoxel pattern analysis and functional connectivity.

**********

Neural Architecture Generator Optimization

Robin Ru, Pedro Esperança, Fabio Maria Carlucci

Neural Architecture Search (NAS) was first proposed to achieve state-of-the-art performance through the discovery of new architecture patterns, without human intervention. An over-reliance on expert knowledge in the search space design has however led to increased performance (local optima) without significant architectural breakthroughs, thus preventing truly novel solutions from being reached. In this work we 1) are the first to investigate casting NAS as a problem of finding the optimal network generator and 2) we propose a new, hierarchical and graph-based search space capable of representing an extremely large variety of network types, yet only requiring few continuous hyper-parameters. This greatly reduces the dimensionality of the problem, enabling the effective use of Bayesian Optimisation as a search strategy. At the same time, we expand the range of valid architectures, motivating a multi-objective learning approach. We demonstrate the effectiveness of this strategy on six benchmark datasets and show that our search space generates extremely lightweight yet highly competitive models.

**********

A Bandit Learning Algorithm and Applications to Auction Design

Kim Thang Nguyen

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

**********
MetaPoison: Practical General-purpose Clean-label Data Poisoning
W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, Tom Goldstein
Data poisoning---the process by which an attacker takes control of a model by making imperceptible changes to a subset of the training data---is an emerging threat in the context of neural networks. Existing attacks for data poisoning neural networks have relied on hand-crafted heuristics, because solving the poisoning problem directly via bilevel optimization is generally thought of as intractable for deep models. We propose MetaPoison, a first-order method that approximates the bilevel problem via meta-learning and crafts poisons that fool neural networks. MetaPoison is effective: it outperforms previous clean-label poisoning methods by a large margin. MetaPoison is robust: poisoned data made for one model transfer to a variety of victim models with unknown training settings and architectures. MetaPoison is general-purpose, it works not only in fine-tuning scenarios, but also for end-to-end training from scratch, which till now hasn't been feasible for clean-label attacks with deep nets. MetaPoison can achieve arbitrary adversary goals---like using poisons of one class to make a target image don the label of another arbitrarily chosen class. Finally, MetaPoison works in the real-world. We demonstrate for the first time successful data poisoning of models trained on the black-box Google Cloud AutoML API.
**********
Sample Efficient Reinforcement Learning via Low-Rank Matrix Estimation
Devavrat Shah, Dogyoon Song, Zhi Xu, Yuzhe 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.
**********
Training Generative Adversarial Networks with Limited Data
Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, Timo Aila
Training generative adversarial networks (GAN) using too little data typically leads to discriminator overfitting, causing training to diverge. We propose an adaptive discriminator augmentation mechanism that significantly stabilizes training in limited data regimes. The approach does not require changes to loss functions or network architectures, and is applicable both when training from scratch and when fine-tuning an existing GAN on another dataset. We demonstrate, on several datasets, that good results are now possible using only a few thousand training images, often matching StyleGAN2 results with an order of magnitude fewer images. We expect this to open up new application domains for GANs. We also find that the widely used CIFAR-10 is, in fact, a limited data benchmark, and improve the record FID from 5.59 to 2.42.
**********
Deeply Learned Spectral Total Variation Decomposition
Tamara G. Grossmann, Yury Korolev, Guy Gilboa, Carola Schoenlieb
Non-linear spectral decompositions of images based on one-homogeneous functionals such as total variation have gained considerable attention in the last few years. Due to their ability to extract spectral components corresponding to objects of different size and contrast, such decompositions enable filtering, feature transfer, image fusion and other applications. However, obtaining this decomposition involves solving multiple non-smooth optimisation problems and is therefore computationally highly intensive. In this paper, we present a neural network approximation of a non-linear spectral decomposition. We report up to four orders of magnitude (×10,000) speedup in processing of mega-pixel size images, compared to classical GPU implementations. Our proposed network, TVspecNET, is able to implicitly learn the underlying PDE and, despite being entirely data driven, inherits invariances of the model based transform. To the best of our knowledge, this is the first approach towards learning a non-linear spectral decomposition of images. Not only do we gain a staggering computational advantage, but this approach can also be seen as a step towards studying neural networks that can decompos

e an image into spectral components defined by a user rather than a handcrafted functional.
**********

FracTrain: Fractionally Squeezing Bit Savings Both Temporally and Spatially for Efficient DNN Training
Yonggan Fu, Haoran You, Yang Zhao, Yue Wang, Chaojian Li, Kailash Gopalakrishnan, Zhangyang Wang, Yingyan Lin
Recent breakthroughs in deep neural networks (DNNs) have fueled a tremendous demand for intelligent edge devices featuring on-site learning, while the practical realization of such systems remains a challenge due to the limited resources available at the edge and the required massive training costs for state-of-the-art (SOTA) DNNs. As reducing precision is one of the most effective knobs for boosting training time/energy efficiency, there has been a growing interest in low-precision DNN training. In this paper, we explore from an orthogonal direction: how to fractionally squeeze out more training cost savings from the most redundant bit level, progressively along the training trajectory and dynamically per input. Specifically, we propose FracTrain that integrates (i) progressive fractional quantization which gradually increases the precision of activations, weights, and gradients that will not reach the precision of SOTA static quantized DNN training until the final training stage, and (ii) dynamic fractional quantization which assigns precisions to both the activations and gradients of each layer in an input-adaptive manner, for only "fractionally" updating layer parameters. Extensive simulations and ablation studies (six models, four datasets, and three training settings including standard, adaptation, and fine-tuning) validate the effectiveness of FracTrain in reducing computational cost and hardware-quantified energy/latency of DNN training while achieving a comparable or better (-0.12%~+1.87%) accuracy. For example, when training ResNet-74 on CIFAR-10, FracTrain achieves 77.6% and 53.5% computational cost and training latency savings, respectively, compared with the best SOTA baseline, while achieving a comparable (-0.07%) accuracy. Our codes are available at: https://github.com/RICE-EIC/FracTrain.
**********

Improving Neural Network Training in Low Dimensional Random Bases
Frithjof Gressmann, Zach Eaton-Rosen, Carlo Luschi
Stochastic Gradient Descent (SGD) has proven to be remarkably effective in optimizing deep neural networks that employ ever-larger numbers of parameters. Yet, improving the efficiency of large-scale optimization remains a vital and highly active area of research. Recent work has shown that deep neural networks can be optimized in randomly-projected subspaces of much smaller dimensionality than their native parameter space. While such training is promising for more efficient and scalable optimization schemes, its practical application is limited by inferior optimization performance.
Here, we improve on recent random subspace approaches as follows. We show that keeping the random projection fixed throughout training is detrimental to optimization. We propose re-drawing the random subspace at each step, which yields significantly better performance. We realize further improvements by applying independent projections to different parts of the network, making the approximation more efficient as network dimensionality grows.
To implement these experiments, we leverage hardware-accelerated pseudo-random number generation to construct the random projections on-demand at every optimization step, allowing us to distribute the computation of independent random directions across multiple workers with shared random seeds. This yields significant reductions in memory and is up to 10x faster for the workloads in question.
**********

Safe Reinforcement Learning via Curriculum Induction
Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, Alekh Agarwal
In safety-critical applications, autonomous agents may need to learn in an environment where mistakes can be very costly. In such settings, the agent needs to behave safely not only after but also while learning. To achieve this, existing safe reinforcement learning methods make an agent rely on priors that let it avoid dangerous situations during exploration with high probability, but both the pr

obabilistic guarantees and the smoothness assumptions inherent in the priors are not viable in many scenarios of interest such as autonomous driving. This paper presents an alternative approach inspired by human teaching, where an agent learns under the supervision of an automatic instructor that saves the agent from violating constraints during learning. In this model, we introduce the monitor that neither needs to know how to do well at the task the agent is learning nor needs to know how the environment works. Instead, it has a library of reset controllers that it activates when the agent starts behaving dangerously, preventing it from doing damage. Crucially, the choices of which reset controller to apply in which situation affect the speed of agent learning. Based on observing agents' progress the teacher itself learns a policy for choosing the reset controllers, a curriculum, to optimize the agent's final policy reward. Our experiments use this framework in two environments to induce curricula for safe and efficient learning.

**********

Leverage the Average: an Analysis of KL Regularization in Reinforcement Learning

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Remi Munos, Matthieu Geist

Recent Reinforcement Learning (RL) algorithms making use of Kullback-Leibler (KL) regularization as a core component have shown outstanding performance. Yet, only little is understood theoretically about why KL regularization helps, so far. We study KL regularization within an approximate value iteration scheme and show that it implicitly averages q-values. Leveraging this insight, we provide a very strong performance bound, the very first to combine two desirable aspects: a linear dependency to the horizon (instead of quadratic) and an error propagation term involving an averaging effect of the estimation errors (instead of an accumulation effect). We also study the more general case of an additional entropy regularizer. The resulting abstract scheme encompasses many existing RL algorithms. Some of our assumptions do not hold with neural networks, so we complement this theoretical analysis with an extensive empirical study.

**********

How Robust are the Estimated Effects of Nonpharmaceutical Interventions against COVID-19?

Mrinank Sharma, Sören Mindermann, Jan Brauner, Gavin Leech, Anna Stephenson, Tomáš Gaven■iak, Jan Kulveit, Yee Whye Teh, Leonid Chindelevitch, Yarin Gal

To what extent are effectiveness estimates of nonpharmaceutical interventions (NPIs) against COVID-19 influenced by the assumptions our models make? To answer this question, we investigate 2 state-of-the-art NPI effectiveness models and propose 6 variants that make different structural assumptions. In particular, we investigate how well NPI effectiveness estimates generalise to unseen countries, and their sensitivity to unobserved factors. Models which account for noise in disease transmission compare favourably. We further evaluate how robust estimates are to different choices of epidemiological parameters and data. Focusing on models that assume transmission noise, we find that previously published results are robust across these choices and across different models. Finally, we mathematically ground the interpretation of NPI effectiveness estimates when certain common assumptions do not hold.

**********

Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses

Kaivalya Rawal, Himabindu Lakkaraju

As predictive models are increasingly being deployed in high-stakes decision-making, there has been a lot of interest in developing algorithms which can provide recourses to affected individuals. While developing such tools is important, it is even more critical to analyze and interpret a predictive model, and vet it thoroughly to ensure that the recourses it offers are meaningful and non-discriminatory before it is deployed in the real world. To this end, we propose a novel model agnostic framework called Actionable Recourse Summaries (AReS) to construct global counterfactual explanations which provide an interpretable and accurate summary of recourses for the entire population. We formulate a novel objective

which simultaneously optimizes for correctness of the recourses and interpretability of the explanations, while minimizing overall recourse costs across the entire population. More specifically, our objective enables us to learn, with optimality guarantees on recourse correctness, a small number of compact rule sets each of which capture recourses for well defined subpopulations within the data. We also demonstrate theoretically that several of the prior approaches proposed to generate recourses for individuals are special cases of our framework. Experimental evaluation with real world datasets and user studies demonstrate that our framework can provide decision makers with a comprehensive overview of recourses corresponding to any black box model, and consequently help detect undesirable model biases and discrimination.

**********

Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization

Benjamin Aubin, Florent Krzakala, Yue Lu, Lenka Zdeborová

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

**********

Projection Efficient Subgradient Method and Optimal Nonsmooth Frank-Wolfe Method

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.

**********

PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks

Minh Vu, My T. Thai

In Graph Neural Networks (GNNs), the graph structure is incorporated into the learning of node representations. This complex structure makes explaining GNNs' predictions become much more challenging. In this paper, we propose PGM-Explainer, a Probabilistic Graphical Model (PGM) model-agnostic explainer for GNNs. Given a prediction to be explained, PGM-Explainer identifies crucial graph components and generates an explanation in form of a PGM approximating that prediction. Different from existing explainers for GNNs where the explanations are drawn from a set of linear functions of explained features, PGM-Explainer is able to demonstrate the dependencies of explained features in form of conditional probabilities. Our theoretical analysis shows that the PGM generated by PGM-Explainer includes the Markov-blanket of the target prediction, i.e. including all its statistical information. We also show that the explanation returned by PGM-Explainer contains the same set of independence statements in the perfect map. Our experiments on both synthetic and real-world datasets show that PGM-Explainer achieves better performance than existing explainers in many benchmark tasks.

**********

Few-Cost Salient Object Detection with Adversarial-Paced Learning

Dingwen Zhang, HaiBin Tian, Jungong Han

Detecting and segmenting salient objects from given image scenes has received great attention in recent years. A fundamental challenge in training the existing deep saliency detection models is the requirement of large amounts of annotated data. While gathering large quantities of training data becomes cheap and easy, annotating the data is an expensive process in terms of time, labor and human expertise. To address this problem, this paper proposes to learn the effective salient object detection model based on the manual annotation on a few training images only, thus dramatically alleviating human labor in training models. To this end, we name this new task as the few-cost salient object detection and propose an adversarial-paced learning (APL)-based framework to facilitate the few-cost learning scenario. Essentially, APL is derived from the self-paced learning (SPL) regime but it infers the robust learning pace through the data-driven adversari

al learning mechanism rather than the heuristic design of the learning regulariz
er. Comprehensive experiments on four widely-used benchmark datasets have demons
trated that the proposed approach can effectively approach to the existing super
vised deep salient object detection models with only 1k human-annotated training
 images.
**********

Minimax Estimation of Conditional Moment Models
Nishanth Dikkala, Greg Lewis, Lester Mackey, Vasilis Syrgkanis
We develop an approach for estimating models described via conditional moment re
strictions, with a prototypical application being non-parametric instrumental va
riable regression. We introduce a min-max criterion function, under which the es
timation problem can be thought of as solving a zero-sum game between a modeler
who is optimizing over the hypothesis space of the target model and an adversary
 who identifies violating moments over a test function space. We analyze the sta
tistical estimation rate of the resulting estimator for arbitrary hypothesis spa
ces, with respect to an appropriate analogue of the mean squared error metric, f
or ill-posed inverse problems. We show that when the minimax criterion is regula
rized with a second moment penalty on the test function and the test function sp
ace is sufficiently rich, then the estimation rate scales with the critical radi
us of the hypothesis and test function spaces, a quantity which typically gives
tight fast rates. Our main result follows from a novel localized Rademacher anal
ysis of statistical learning problems defined via minimax objectives. We provide
 applications of our main results for several hypothesis spaces used in practice
 such as: reproducing kernel Hilbert spaces, high dimensional sparse linear func
tions, spaces defined via shape constraints, ensemble estimators such as random
forests, and neural networks. For each of these applications we provide computat
ionally efficient optimization methods for solving the corresponding minimax pro
blem (e.g. stochastic first-order heuristics for neural networks). In several ap
plications, we show how our modified mean squared error rate, combined with cond
itions that bound the ill-posedness of the inverse problem, lead to mean squared
 error rates. We conclude with an extensive experimental analysis of the propose
d methods.
**********

Causal Imitation Learning With Unobserved Confounders
Junzhe Zhang, Daniel Kumor, Elias Bareinboim
One of the common ways children learn is by mimicking adults. Imitation learning
 focuses on learning policies with suitable performance from demonstrations gene
rated by an expert, with an unspecified performance measure, and unobserved rewa
rd signal. Popular methods for imitation learning start by either directly mimic
king the behavior policy of an expert (behavior cloning) or by learning a reward
 function that prioritizes observed expert trajectories (inverse reinforcement l
earning). However, these methods rely on the assumption that covariates used by
the expert to determine her/his actions are fully observed. In this paper, we re
lax this assumption and study imitation learning when sensory inputs of the lear
ner and the expert differ. First, we provide a non-parametric, graphical criteri
on that is complete (both necessary and sufficient) for determining the feasibil
ity of imitation from the combinations of demonstration data and qualitative ass
umptions about the underlying environment, represented in the form of a causal m
odel. We then show that when such a criterion does not hold, imitation could sti
ll be feasible by exploiting quantitative knowledge of the expert trajectories.
Finally, we develop an efficient procedure for learning the imitating policy fro
m experts' trajectories.
**********

Your GAN is Secretly an Energy-based Model and You Should Use Discriminator Driv
en Latent Sampling
Tong Che, Ruixiang ZHANG, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yu
an Cao, Yoshua Bengio
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth

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

Learning Black-Box Attackers with Transferable Priors and Query Feedback
Jiancheng YANG, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, Chenglong Zhao
This paper addresses the challenging black-box adversarial attack problem, where only classification confidence of a victim model is available. Inspired by consistency of visual saliency between different vision models, a surrogate model is expected to improve the attack performance via transferability. By combining transferability-based and query-based black-box attack, we propose a surprisingly simple baseline approach (named SimBA++) using the surrogate model, which significantly outperforms several state-of-the-art methods. Moreover, to efficiently utilize the query feedback, we update the surrogate model in a novel learning scheme, named High-Order Gradient Approximation (HOGA). By constructing a high-order gradient computation graph, we update the surrogate model to approximate the victim model in both forward and backward pass. The SimBA++ and HOGA result in Learnable Black-Box Attack (LeBA), which surpasses previous state of the art by considerable margins: the proposed LeBA significantly reduces queries, while keeping higher attack success rates close to 100% in extensive ImageNet experiments, including attacking vision benchmarks and defensive models. Code is open source at https://github.com/TrustworthyDL/LeBA.
**********

Locally Differentially Private (Contextual) Bandits Learning
Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, 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-authors prior to requesting a name change in the electronic proceedings.
**********

Invertible Gaussian Reparameterization: Revisiting the Gumbel-Softmax
Andres Potapczynski, Gabriel Loaiza-Ganem, John P. Cunningham
The Gumbel-Softmax is a continuous distribution over the simplex that is often used as a relaxation of discrete distributions. Because it can be readily interpreted and easily reparameterized, it enjoys widespread use. We propose a modular and more flexible family of reparameterizable distributions where Gaussian noise is transformed into a one-hot approximation through an invertible function. This invertible function is composed of a modified softmax and can incorporate diverse transformations that serve different specific purposes. For example, the stick-breaking procedure allows us to extend the reparameterization trick to distributions with countably infinite support, thus enabling the use of our distribution along nonparametric models, or normalizing flows let us increase the flexibility of the distribution. Our construction enjoys theoretical advantages over the Gumbel-Softmax, such as closed form KL, and significantly outperforms it in a variety of experiments. Our code is available at https://github.com/cunningham-lab/igr.
**********

Kernel Based Progressive Distillation for Adder Neural Networks
Yixing Xu, Chang Xu, Xinghao Chen, Wei Zhang, Chunjing XU, Yunhe 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.
**********

Adversarial Soft Advantage Fitting: Imitation Learning without Policy Optimization
Paul Barde, Julien Roy, Wonseok Jeon, Joelle Pineau, Chris Pal, Derek Nowrouzezahrai
Adversarial Imitation Learning alternates between learning a discriminator -- which tells apart expert's demonstrations from generated ones -- and a generator's policy to produce trajectories that can fool this discriminator. This alternated optimization is known to be delicate in practice since it compounds unstable a

dversarial training with brittle and sample-inefficient reinforcement learning. We propose to remove the burden of the policy optimization steps by leveraging a novel discriminator formulation. Specifically, our discriminator is explicitly conditioned on two policies: the one from the previous generator's iteration and a learnable policy. When optimized, this discriminator directly learns the optimal generator's policy. Consequently, our discriminator's update solves the generator's optimization problem for free: learning a policy that imitates the expert does not require an additional optimization loop. This formulation effectively cuts by half the implementation and computational burden of Adversarial Imitation Learning algorithms by removing the Reinforcement Learning phase altogether. We show on a variety of tasks that our simpler approach is competitive to prevalent Imitation Learning methods.

**********

Agree to Disagree: Adaptive Ensemble Knowledge Distillation in Gradient Space
Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, Changshui Zhang

Distilling knowledge from an ensemble of teacher models is expected to have a more promising performance than that from a single one. Current methods mainly adopt a vanilla average rule, i.e., to simply take the average of all teacher losses for training the student network. However, this approach treats teachers equally and ignores the diversity among them. When conflicts or competitions exist among teachers, which is common, the inner compromise might hurt the distillation performance. In this paper, we examine the diversity of teacher models in the gradient space and regard the ensemble knowledge distillation as a multi-objective optimization problem so that we can determine a better optimization direction for the training of student network. Besides, we also introduce a tolerance parameter to accommodate disagreement among teachers. In this way, our method can be seen as a dynamic weighting method for each teacher in the ensemble. Extensive experiments validate the effectiveness of our method for both logits-based and feature-based cases.

**********

The Wasserstein Proximal Gradient Algorithm
Adil Salim, Anna Korba, Giulia Luise

Wasserstein gradient flows are continuous time dynamics that define curves of steepest descent to minimize an objective function over the space of probability measures (i.e., the Wasserstein space). This objective is typically a divergence w.r.t. a fixed target distribution. In recent years, these continuous time dynamics have been used to study the convergence of machine learning algorithms aiming at approximating a probability distribution. However, the discrete-time behavior of these algorithms might differ from the continuous time dynamics. Besides, although discretized gradient flows have been proposed in the literature, little is known about their minimization power. In this work, we propose a Forward Backward (FB) discretization scheme that can tackle the case where the objective function is the sum of a smooth and a nonsmooth geodesically convex terms. Using techniques from convex optimization and optimal transport, we analyze the FB scheme as a minimization algorithm on the Wasserstein space. More precisely, we show under mild assumptions that the FB scheme has convergence guarantees similar to the proximal gradient algorithm in Euclidean spaces (resp. similar to the associated Wasserstein gradient flow).

**********

Universally Quantized Neural Compression
Eirikur Agustsson, Lucas Theis

A popular approach to learning encoders for lossy compression is to use additive uniform noise during training as a differentiable approximation to test-time quantization. We demonstrate that a uniform noise channel can also be implemented at test time using universal quantization (Ziv, 1985). This allows us to eliminate the mismatch between training and test phases while maintaining a completely differentiable loss function. Implementing the uniform noise channel is a special case of the more general problem of communicating a sample, which we prove is computationally hard if we do not make assumptions about its distribution. Howev

er, the uniform special case is efficient as well as easy to implement and thus of great interest from a practical point of view. Finally, we show that quantization can be obtained as a limiting case of a soft quantizer applied to the uniform noise channel, bridging compression with and without quantization.
**********

Temporal Variability in Implicit Online Learning
Nicolò Campolongo, Francesco Orabona
In the setting of online learning, Implicit algorithms turn out to be highly successful from a practical standpoint. However, the tightest regret analyses only show marginal improvements over Online Mirror Descent. In this work, we shed light on this behavior carrying out a careful regret analysis. We prove a novel static regret bound that depends on the temporal variability of the sequence of loss functions, a quantity which is often encountered when considering dynamic comp etitors. We show, for example, that the regret can be constant if the temporal variability is constant and the learning rate is tuned appropriately, without the need of smooth losses. Moreover, we present an adaptive algorithm that achieves this regret bound without prior knowledge of the temporal variability and prove a matching lower bound. Finally, we validate our theoretical findings on classification and regression datasets.
**********

Investigating Gender Bias in Language Models Using Causal Mediation Analysis
Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, Stuart Shieber
Many interpretation methods for neural models in natural language processing investigate how information is encoded inside hidden representations. However, these methods can only measure whether the information exists, not whether it is actually used by the model.  We propose a methodology grounded in the theory of causal mediation analysis for interpreting which parts of a model are causally implicated in its behavior. The approach enables us to analyze the mechanisms that facilitate the flow of information from input to output through various model components, known as mediators. As a case study, we apply this methodology to analyzing gender bias in pre-trained Transformer language models. We study the role of individual neurons and attention heads in mediating gender bias across three datasets designed to gauge a model's sensitivity to gender bias. Our mediation analysis reveals that gender bias effects are concentrated in specific components of the model that may exhibit highly specialized behavior.
**********

Off-Policy Imitation Learning from Observations
Zhuangdi Zhu, Kaixiang Lin, Bo Dai, Jiayu Zhou
Learning from Observations (LfO) is a practical reinforcement learning scenario from which many applications can benefit through the reuse of incomplete resources. Compared to conventional imitation learning (IL), LfO is more challenging because of the lack of expert action guidance.
In both conventional IL and LfO, distribution matching is at the heart of their foundation. Traditional distribution matching approaches are sample-costly which depend on on-policy transitions for policy learning. Towards sample-efficiency, some off-policy solutions have been proposed, which, however, either lack comprehensive theoretical justifications or depend on the guidance of expert actions. In this work, we propose a sample-efficient LfO approach which enables off-policy optimization in a principled manner. To further accelerate the learning procedure, we regulate the policy update with an inverse action model, which assists distribution matching from the perspective of mode-covering. Extensive empirical results on challenging locomotion tasks indicate that our approach is comparable with state-of-the-art in terms of both sample-efficiency and asymptotic performance.
**********

Escaping Saddle-Point Faster under Interpolation-like Conditions
Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, Prasant Mohapatra
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.

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

**********

Matérn Gaussian Processes on Riemannian Manifolds

Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, Marc Deisenroth (he/him)

Gaussian processes are an effective model class for learning unknown functions, particularly in settings where accurately representing predictive uncertainty is of key importance. Motivated by applications in the physical sciences, the widely-used Matérn class of Gaussian processes has recently been generalized to model functions whose domains are Riemannian manifolds, by re-expressing said processes as solutions of stochastic partial differential equations. In this work, we propose techniques for computing the kernels of these processes on compact Riemannian manifolds via spectral theory of the Laplace-Beltrami operator in a fully constructive manner, thereby allowing them to be trained via standard scalable techniques such as inducing point methods. We also extend the generalization from the Matérn to the widely-used squared exponential Gaussian process. By allowing Riemannian Matérn Gaussian processes to be trained using well-understood techniques, our work enables their use in mini-batch, online, and non-conjugate settings, and makes them more accessible to machine learning practitioners.

**********

Improved Techniques for Training Score-Based Generative Models

Yang Song, Stefano Ermon

Score-based generative models can produce high quality image samples comparable to GANs, without requiring adversarial optimization. However, existing training procedures are limited to images of low resolution (typically below 32 x 32), and can be unstable under some settings. We provide a new theoretical analysis of learning and sampling from score models in high dimensional spaces, explaining existing failure modes and motivating new solutions that generalize across datasets.

To enhance stability, we also propose to maintain an exponential moving average of model weights. With these improvements, we can effortlessly scale score-based generative models to images with unprecedented resolutions ranging from 64 x 64 to 256 x 256. Our score-based models can generate high-fidelity samples that rival best-in-class GANs on various image datasets, including CelebA, FFHQ, and multiple LSUN categories.

**********

wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, Michael Auli

We show for the first time that learning powerful representations from speech audio alone followed by fine-tuning on transcribed speech can outperform the best semi-supervised methods while being conceptually simpler. wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned. Experiments using all labeled data of Librispeech achieve 1.8/3.3 WER on the clean/other test sets. When lowering the amount of labeled data to one hour, wav2vec 2.0 outperforms the previous state of the art on the 100 hour subset while using 100 times less labeled data. Using just ten minutes of labeled data and pre-training on 53k hours of unlabeled data still achieves 4.8/8.2 WER. This demonstrates the feasibility of speech recognition with limited amounts of labeled data.

**********

A Maximum-Entropy Approach to Off-Policy Evaluation in Average-Reward MDPs

Nevena Lazic, Dong Yin, Mehrdad Farajtabar, Nir Levine, Dilan Gorur, Chris Harris, Dale Schuurmans

This work focuses on off-policy evaluation (OPE) with function approximation in infinite-horizon undiscounted Markov decision processes (MDPs). For MDPs that are ergodic and linear (i.e. where rewards and dynamics are linear in some known features), we provide the first finite-sample OPE error bound, extending the existing results beyond the episodic and discounted cases. In a more general setting, when the feature dynamics are approximately linear and for arbitrary rewards,

we propose a new approach for estimating stationary distributions with function approximation. We formulate this problem as finding the maximum-entropy distribution subject to matching feature expectations under empirical dynamics. We show that this results in an exponential-family distribution whose sufficient statistics are the features, paralleling maximum-entropy approaches in supervised learning. We demonstrate the effectiveness of the proposed OPE approaches in multiple environments.

**********

Instead of Rewriting Foreign Code for Machine Learning, Automatically Synthesize Fast Gradients

William Moses, Valentin Churavy

Applying differentiable programming techniques and machine learning algorithms to foreign programs requires developers to either rewrite their code in a machine learning framework, or otherwise provide derivatives of the foreign code. This paper presents Enzyme, a high-performance automatic differentiation (AD) compiler plugin for the LLVM compiler framework capable of synthesizing gradients of statically analyzable programs expressed in the LLVM intermediate representation (IR). Enzyme synthesizes gradients for programs written in any language whose compiler targets LLVM IR including C, C++, Fortran, Julia, Rust, Swift, MLIR, etc., thereby providing native AD capabilities in these languages. Unlike traditional source-to-source and operator-overloading tools, Enzyme performs AD on optimized IR. On a machine-learning focused benchmark suite including Microsoft's ADBench, AD on optimized IR achieves a geometric mean speedup of 4.2 times over AD on IR before optimization allowing Enzyme to achieve state-of-the-art performance. Packaging Enzyme for PyTorch and TensorFlow provides convenient access to gradients of foreign code with state-of-the-art performance, enabling foreign code to be directly incorporated into existing machine learning workflows.

**********

Does Unsupervised Architecture Representation Learning Help Neural Architecture Search?

Shen Yan, Yu Zheng, Wei Ao, Xiao Zeng, Mi Zhang

Existing Neural Architecture Search (NAS) methods either encode neural architectures using discrete encodings that do not scale well, or adopt supervised learning-based methods to jointly learn architecture representations and optimize architecture search on such representations which incurs search bias. Despite the widespread use, architecture representations learned in NAS are still poorly understood. We observe that the structural properties of neural architectures are hard to preserve in the latent space if architecture representation learning and search are coupled, resulting in less effective search performance. In this work, we find empirically that pre-training architecture representations using only neural architectures without their accuracies as labels improves the downstream architecture search efficiency. To explain this finding, we visualize how unsupervised architecture representation learning better encourages neural architectures with similar connections and operators to cluster together. This helps map neural architectures with similar performance to the same regions in the latent space and makes the transition of architectures in the latent space relatively smooth, which considerably benefits diverse downstream search strategies.

**********

Value-driven Hindsight Modelling

Arthur Guez, Fabio Viola, Theophane Weber, Lars Buesing, Steven Kapturowski, Doina Precup, David Silver, Nicolas Heess

Value estimation is a critical component of the reinforcement learning (RL) paradigm. The question of how to effectively learn value predictors from data is one of the major problems studied by the RL community, and different approaches exploit structure in the problem domain in different ways. Model learning can make use of the rich transition structure present in sequences of observations, but this approach is usually not sensitive to the reward function. In contrast, model-free methods directly leverage the quantity of interest from the future, but receive a potentially weak scalar signal (an estimate of the return). We develop an approach for representation learning in RL that sits in between these two ext

remes: we propose to learn what to model in a way that can directly help value p
rediction. To this end, we determine which features of the future trajectory pro
vide useful information to predict the associated return. This provides tractabl
e prediction targets that are directly relevant for a task, and can thus acceler
ate learning the value function. The idea can be understood as reasoning, in hin
dsight, about which aspects of the future observations could help past value pre
diction. We show how this can help dramatically even in simple policy evaluation
 settings. We then test our approach at scale in challenging domains, including
on 57 Atari 2600 games.
**********

Dynamic Regret of Convex and Smooth Functions
Peng Zhao, Yu-Jie Zhang, Lijun Zhang, Zhi-Hua Zhou
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
**********

On Convergence of Nearest Neighbor Classifiers over Feature Transformations
Luka Rimanic, Cedric Renggli, Bo Li, Ce Zhang
The k-Nearest Neighbors (kNN) classifier is a fundamental non-parametric machine
 learning algorithm. However, it is well known that it suffers from the curse of
 dimensionality, which is why in practice one often applies a kNN classifier on
top of a (pre-trained) feature transformation. From a theoretical perspective, m
ost, if not all theoretical results aimed at understanding the kNN classifier ar
e derived for the raw feature space. This leads to an emerging gap between our t
heoretical understanding of kNN and its practical applications.
In this paper, we take a first step towards bridging this gap. We provide a nove
l analysis on the convergence rates of a kNN classifier over transformed feature
s. This analysis requires in-depth understanding of the properties that connect
both the transformed space and the raw feature space. More precisely, we build o
ur convergence bound upon two key properties of the transformed space: (1) safet
y -- how well can one recover the raw posterior from the transformed space, and
(2) smoothness -- how complex this recovery function is. Based on our result, we
  are able to explain why some (pre-trained) feature transformations are better
suited for a kNN classifier than other. We empirically validate that both proper
ties have an impact on the kNN convergence on 30 feature transformations with 6
benchmark datasets spanning from the vision to the text domain.
**********

Mitigating Manipulation in Peer Review via Randomized Reviewer Assignments
Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar Shah, Vincent Conitzer, Fei Fang
We consider three important challenges in conference peer review: (i) reviewers
maliciously attempting to get assigned to certain papers to provide positive rev
iews, possibly as part of quid-pro-quo arrangements with the authors; (ii) "torp
edo reviewing," where reviewers deliberately attempt to get assigned to certain
papers that they dislike in order to reject them; (iii) reviewer de-anonymizatio
n on release of the similarities and the reviewer-assignment code. On the concep
tual front, we identify connections between these three problems and present a f
ramework that brings all these challenges under a common umbrella. We then prese
nt a (randomized) algorithm for reviewer assignment that can optimally solve the
 reviewer-assignment problem under any given constraints on the probability of a
ssignment for any reviewer-paper pair. We further consider the problem of restri
cting the joint probability that certain suspect pairs of reviewers are assigned
 to certain papers, and show that this problem is NP-hard for arbitrary constrai
nts on these joint probabilities but efficiently solvable for a practical specia
l case. Finally, we experimentally evaluate our algorithms on datasets from past
 conferences, where we observe that they can limit the chance that any malicious
 reviewer gets assigned to their desired paper to 50% while producing assignment
s with over 90% of the total optimal similarity.
**********

Contrastive learning of global and local features for medical image segmentation

with limited annotations
Krishna Chaitanya, Ertunc Erdil, Neerav Karani, Ender Konukoglu
A key requirement for the success of supervised deep learning is a large labeled dataset - a condition that is difficult to meet in medical image analysis. Self-supervised learning (SSL) can help in this regard by providing a strategy to pre-train a neural network with unlabeled data, followed by fine-tuning for a downstream task with limited annotations. Contrastive learning, a particular variant of SSL, is a powerful technique for learning image-level representations. In this work, we propose strategies for extending the contrastive learning framework for segmentation of volumetric medical images in the semi-supervised setting with limited annotations, by leveraging domain-specific and problem-specific cues. Specifically, we propose (1) novel contrasting strategies that leverage structural similarity across volumetric medical images (domain-specific cue) and (2) a local version of the contrastive loss to learn distinctive representations of local regions that are useful for per-pixel segmentation (problem-specific cue). We carry out an extensive evaluation on three Magnetic Resonance Imaging (MRI) datasets. In the limited annotation setting, the proposed method yields substantial improvements compared to other self-supervision and semi-supervised learning techniques. When combined with a simple data augmentation technique, the proposed method reaches within 8\% of benchmark performance using only two labeled MRI volumes for training. The code is made public at https://github.com/krishnabits001/domain_specific_cl.
**********
Self-Supervised Graph Transformer on Large-Scale Molecular Data
Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying WEI, Wenbing Huang, Junzhou Huang
How to obtain informative representations of molecules is a crucial prerequisite in AI-driven drug design and discovery. Recent researches abstract molecules as graphs and employ Graph Neural Networks (GNNs) for molecular representation learning. Nevertheless, two issues impede the usage of GNNs in real scenarios: (1) insufficient labeled molecules for supervised training; (2) poor generalization capability to new-synthesized molecules. To address them both, we propose a novel framework, GROVER, which stands for Graph Representation frOm self-superVised mEssage passing tRansformer. With carefully designed self-supervised tasks in node-, edge- and graph-level, GROVER can learn rich structural and semantic information of molecules from enormous unlabelled molecular data. Rather, to encode such complex information, GROVER integrates Message Passing Networks into the Transformer-style architecture to deliver a class of more expressive encoders of molecules. The flexibility of GROVER allows it to be trained efficiently on large-scale molecular dataset without requiring any supervision, thus being immunized to the two issues mentioned above. We pre-train GROVER with 100 million parameters on 10 million unlabelled molecules---the biggest GNN and the largest training dataset in molecular representation learning. We then leverage the pre-trained GROVER for molecular property prediction followed by task-specific fine-tuning, where we observe a huge improvement (more than 6% on average) from current state-of-the-art methods on 11 challenging benchmarks. The insights we gained are that well-designed self-supervision losses and largely-expressive pre-trained models enjoy the significant potential on performance boosting.
**********
Generative Neurosymbolic Machines
Jindong Jiang, Sungjin Ahn
Reconciling symbolic and distributed representations is a crucial challenge that can potentially resolve the limitations of current deep learning. Remarkable advances in this direction have been achieved recently via generative object-centric representation models. While learning a recognition model that infers object-centric symbolic representations like bounding boxes from raw images in an unsupervised way, no such model can provide another important ability of a generative model, i.e., generating (sampling) according to the structure of learned world density. In this paper, we propose Generative Neurosymbolic Machines, a generative model that combines the benefits of distributed and symbolic representations

to support both structured representations of symbolic components and density-ba
sed generation. These two crucial properties are achieved by a two-layer latent
hierarchy with the global distributed latent for flexible density modeling and t
he structured symbolic latent map. To increase the model flexibility in this hie
rarchical structure, we also propose the StructDRAW prior. In experiments, we sh
ow that the proposed model significantly outperforms the previous structured rep
resentation models as well as the state-of-the-art non-structured generative mod
els in terms of both structure accuracy and image generation quality.
**********

How many samples is a good initial point worth in Low-rank Matrix Recovery?
Jialun Zhang, Richard Zhang
Given a sufficiently large amount of labeled data, the nonconvex low-rank matrix
 recovery problem contains no spurious local minima, so a local optimization alg
orithm is guaranteed to converge to a global minimum starting from any initial g
uess. However, the actual amount of data needed by this theoretical guarantee is
 very pessimistic, as it must prevent spurious local minima from existing anywhe
re, including at adversarial locations. In contrast, prior work based on good in
itial guesses have more realistic data requirements, because they allow spurious
 local minima to exist outside of a neighborhood of the solution. In this paper,
 we quantify the relationship between the quality of the initial guess and the c
orresponding reduction in data requirements. Using the restricted isometry const
ant as a surrogate for sample complexity, we compute a sharp “threshold” number
of samples needed to prevent each specific point on the optimization landscape f
rom becoming a spurious local minima. Optimizing the threshold over regions of t
he landscape, we see that, for initial points not too close to the ground truth,
 a linear improvement in the quality of the initial guess amounts to a constant
factor improvement in the sample complexity.
**********

CSER: Communication-efficient SGD with Error Reset
Cong Xie, Shuai Zheng, Sanmi Koyejo, Indranil Gupta, Mu Li, Haibin 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.
**********

Efficient estimation of neural tuning during naturalistic behavior
Edoardo Balzani, Kaushik  Lakshminarasimhan, Dora Angelaki, Cristina Savin
Recent technological advances in systems neuroscience have led to a shift away f
rom using simple tasks, with low-dimensional, well-controlled stimuli, towards t
rying to understand neural activity during naturalistic behavior. However, with
the increase in number and complexity of task-relevant features, standard analys
es such as estimating tuning functions become challenging. Here, we use a Poisso
n generalized additive model (P-GAM) with spline nonlinearities and an exponenti
al link function to map a large number of task variables (input stimuli, behavio
ral outputs, or activity of other neurons, modeled as discrete events or continu
ous variables) into spike counts. We develop efficient procedures for parameter
learning by optimizing a generalized cross-validation score and infer marginal c
onfidence bounds for the contribution of each feature to neural responses. This
allows us to robustly identify a minimal set of task features that each neuron i
s responsive to, circumventing computationally demanding model comparison. We sh
ow that our estimation procedure outperforms traditional regularized GLMs in ter
ms of both fit quality and computing time. When applied to neural recordings fro
m monkeys performing a virtual reality spatial navigation task, P-GAM reveals mi
xed selectivity and preferential coupling between neurons with similar tuning.
**********

High-recall causal discovery for autocorrelated time series with latent confound
ers
Andreas Gerhardus, Jakob Runge
We present a new method for linear and nonlinear, lagged and contemporaneous con
straint-based causal discovery from observational time series in the presence of

latent confounders. We show that existing causal discovery methods such as FCI and variants suffer from low recall in the autocorrelated time series case and identify low effect size of conditional independence tests as the main reason. Information-theoretical arguments show that effect size can often be increased if causal parents are included in the conditioning sets. To identify parents early on, we suggest an iterative procedure that utilizes novel orientation rules to determine ancestral relationships already during the edge removal phase. We prove that the method is order-independent, and sound and complete in the oracle case. Extensive simulation studies for different numbers of variables, time lags, sample sizes, and further cases demonstrate that our method indeed achieves much higher recall than existing methods for the case of autocorrelated continuous variables while keeping false positives at the desired level. This performance gain grows with stronger autocorrelation. At github.com/jakobrunge/tigramite we provide Python code for all methods involved in the simulation studies.
**********

Forget About the LiDAR: Self-Supervised Depth Estimators with MED Probability Volumes
Juan Luis GonzalezBello, Munchurl Kim
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Joint Contrastive Learning with Infinite Possibilities
Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, Tao Mei
This paper explores useful modifications of the recent development in contrastive learning via novel probabilistic modeling. We derive a particular form of contrastive loss named Joint Contrastive Learning (JCL). JCL implicitly involves the simultaneous learning of an infinite number of query-key pairs, which poses tighter constraints when searching for invariant features. We derive an upper bound on this formulation that allows analytical solutions in an end-to-end training manner. While JCL is practically effective in numerous computer vision applications, we also theoretically unveil the certain mechanisms that govern the behavior of JCL. We demonstrate that the proposed formulation harbors an innate agency that strongly favors similarity within each instance-specific class, and therefore remains advantageous when searching for discriminative features among distinct instances. We evaluate these proposals on multiple benchmarks, demonstrating considerable improvements over existing algorithms. Code is publicly available at: https://github.com/caiqi/Joint-Contrastive-Learning.
**********

Robust Gaussian Covariance Estimation in Nearly-Matrix Multiplication Time
Jerry Li, Guanghao Ye
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Adversarially-learned Inference via an Ensemble of Discrete Undirected Graphical Models
Adarsh Keshav Jeewajee, Leslie Kaelbling
Undirected graphical models are compact representations of joint probability distributions over random variables. To solve inference tasks of interest, graphical models of arbitrary topology can be trained using empirical risk minimization. However, to solve inference tasks that were not seen during training, these models (EGMs) often need to be re-trained. Instead, we propose an inference-agnostic adversarial training framework which produces an infinitely-large ensemble of graphical models (AGMs). The ensemble is optimized to generate data within the GAN framework, and inference is performed using a finite subset of these models. AGMs perform comparably with EGMs on inference tasks that the latter were specifically optimized for. Most importantly, AGMs show significantly better generaliz

ation to unseen inference tasks compared to EGMs, as well as deep neural architectures like GibbsNet and VAEAC which allow arbitrary conditioning. Finally, AGMs allow fast data sampling, competitive with Gibbs sampling from EGMs.
**********

GS-WGAN: A Gradient-Sanitized Approach for Learning Differentially Private Generators
Dingfan Chen, Tribhuvanesh Orekondy, Mario Fritz
The wide-spread availability of rich data has fueled the growth of machine learning applications in numerous domains. However, growth in domains with highly-sensitive data (e.g., medical) is largely hindered as the private nature of data prohibits it from being shared. To this end, we propose Gradient-sanitized Wasserstein Generative Adversarial Networks (GS-WGAN), which allows releasing a sanitized form of the sensitive data with rigorous privacy guarantees.
In contrast to prior work, our approach is able to distort gradient information more precisely, and thereby enabling training deeper models which generate more informative samples. Moreover, our formulation naturally allows for training GANs in both centralized and federated (i.e., decentralized) data scenarios.
Through extensive experiments, we find our approach consistently outperforms state-of-the-art approaches across multiple metrics (e.g., sample quality) and datasets.
**********

SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows
Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, Max Welling
Normalizing flows and variational autoencoders are powerful generative models that can represent complicated density functions. However, they both impose constraints on the models: Normalizing flows use bijective transformations to model densities whereas VAEs learn stochastic transformations that are non-invertible and thus typically do not provide tractable estimates of the marginal likelihood.
In this paper, we introduce SurVAE Flows: A modular framework of composable transformations that encompasses VAEs and normalizing flows. SurVAE Flows bridge the gap between normalizing flows and VAEs with surjective transformations, wherein the transformations are deterministic in one direction -- thereby allowing exact likelihood computation, and stochastic in the reverse direction -- hence providing a lower bound on the corresponding likelihood. We show that several recently proposed methods, including dequantization and augmented normalizing flows, can be expressed as SurVAE Flows. Finally, we introduce common operations such as the max value, the absolute value, sorting and stochastic permutation as composable layers in SurVAE Flows.
**********

Learning Causal Effects via Weighted Empirical Risk Minimization
Yonghan Jung, Jin Tian, Elias Bareinboim
Learning causal effects from data is a fundamental problem across the sciences. Determining the identifiability of a target effect from a combination of the observational distribution and the causal graph underlying a phenomenon is well-understood in theory. However, in practice, it remains a challenge to apply the identification theory to estimate the identified causal functionals from finite samples. Although a plethora of effective estimators have been developed under the setting known as the back-door (also called conditional ignorability), there exists still no systematic way of estimating arbitrary causal functionals that are both computationally and statistically attractive. This paper aims to bridge this gap, from causal identification to causal estimation. We note that estimating functionals from limited samples based on the empirical risk minimization (ERM) principle has been pervasive in the machine learning literature, and these methods have been extended to causal inference under the back-door setting. In this paper, we develop a learning framework that marries two families of methods, benefiting from the generality of the causal identification theory and the effectiveness of the estimators produced based on the principle of ERM. Specifically, we develop a sound and complete algorithm that generates causal functionals in the form of weighted distributions that are amenable to the ERM optimization. We then provide a practical procedure for learning causal effects from finite samples

and a causal graph. Finally,  experimental results support the effectiveness of our approach.
**********
Revisiting the Sample Complexity of Sparse Spectrum Approximation of Gaussian Processes
Minh Hoang, Nghia Hoang, Hai Pham, David Woodruff
We introduce a new scalable approximation for Gaussian processes with provable guarantees which holds simultaneously over its entire parameter space. Our approximation is obtained from an improved sample complexity analysis for sparse spectrum Gaussian processes (SSGPs). In particular, our analysis shows that under a certain data disentangling condition, an SSGP's prediction and model evidence (for training) can well-approximate those of a full GP with low sample complexity. We also develop a new auto-encoding algorithm that finds a latent space to disentangle latent input coordinates into well-separated clusters, which is amenable to our sample complexity analysis. We validate our proposed method on several benchmarks with promising results supporting our theoretical analysis.
**********
Incorporating Interpretable Output Constraints in Bayesian Neural Networks
Wanqian Yang, Lars Lorch, Moritz Graule, Himabindu Lakkaraju, Finale Doshi-Velez
Domains where supervised models are deployed often come with task-specific constraints, such as prior expert knowledge on the ground-truth function, or desiderata like safety and fairness. We introduce a novel probabilistic framework for reasoning with such constraints and formulate a prior that enables us to effectively incorporate them into Bayesian neural networks (BNNs), including a variant that can be amortized over tasks. The resulting Output-Constrained BNN (OC-BNN) is fully consistent with the Bayesian framework for uncertainty quantification and is amenable to black-box inference. Unlike typical BNN inference in uninterpretable parameter space, OC-BNNs widen the range of functional knowledge that can be incorporated, especially for model users without expertise in machine learning. We demonstrate the efficacy of OC-BNNs on real-world datasets, spanning multiple domains such as healthcare, criminal justice, and credit scoring.
**********
Multi-Stage Influence Function
Hongge  Chen, Si Si, Yang Li, Ciprian Chelba, Sanjiv Kumar, Duane Boning, Cho-Jui Hsieh
Multi-stage training and knowledge transfer, from a large-scale pretraining task to various finetuning tasks, have revolutionized natural language processing and computer vision resulting in state-of-the-art performance improvements. In this paper,  we develop a  multi-stage influence function score to track predictions from a finetuned model all the way back to the pretraining data. With this score, we can identify the pretraining examples in the pretraining task that contribute most to a prediction in the finetuning task.  The proposed multi-stage influence function generalizes the original influence function for a single model in (Koh &Liang, 2017), thereby enabling influence computation through both pretrained and finetuned models. We study two different scenarios with the pretrained embedding fixed or updated in the finetuning tasks. We test our proposed method in various experiments to show its effectiveness and potential applications.
**********
Probabilistic Fair Clustering
Seyed  Esmaeili, Brian Brubach, Leonidas Tsepenekas, John Dickerson
In clustering problems, a central decision-maker is given a complete metric graph over vertices and must provide a clustering of vertices that minimizes some objective function.  In fair clustering problems, vertices are endowed with a color (e.g., membership in a group), and the requirements of a valid clustering might also include the representation of colors in the solution.  Prior work in fair clustering assumes complete knowledge of group membership. In this paper, we generalize this by assuming imperfect knowledge of group membership through probabilistic assignments, and present algorithms in this more general setting with approximation ratio guarantees.  We also address the problem of "metric membership", where group membership has a notion of order and distance. Experiments are co

nducted using our proposed algorithms as well as baselines to validate our appro
ach, and also surface nuanced concerns when group membership is not known determ
inistically.
**********

Stochastic Segmentation Networks: Modelling Spatially Correlated Aleatoric Uncer
tainty
Miguel Monteiro, Loic Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernar
do Marques, Konstantinos Kamnitsas, Mark van der Wilk, Ben Glocker
In image segmentation, there is often more than one plausible solution for a giv
en input. In medical imaging, for example, experts will often disagree about the
 exact location of object boundaries. Estimating this inherent uncertainty and p
redicting multiple plausible hypotheses is of great interest in many application
s, yet this ability is lacking in most current deep learning methods. In this pa
per, we introduce stochastic segmentation networks (SSNs), an efficient probabil
istic method for modelling aleatoric uncertainty with any image segmentation net
work architecture. In contrast to approaches that produce pixel-wise estimates,
SSNs model joint distributions over entire label maps and thus can generate mult
iple spatially coherent hypotheses for a single image. By using a low-rank multi
variate normal distribution over the logit space to model the probability of the
 label map given the image, we obtain a spatially consistent probability distrib
ution that can be efficiently computed by a neural network without any changes t
o the underlying architecture. We tested our method on the segmentation of real-
world medical data, including lung nodules in 2D CT and brain tumours in 3D mult
imodal MRI scans. SSNs outperform state-of-the-art for modelling correlated unce
rtainty in ambiguous images while being much simpler, more flexible, and more ef
ficient.
**********

ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear I
CA
Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, Aapo Hyvarinen
We consider the identifiability theory of probabilistic models and establish suf
ficient conditions under which the representations learnt by a very broad family
 of conditional energy-based models are unique in function space, up to a simple
 transformation. In our model family, the energy function is the dot-product bet
ween two feature extractors, one for the dependent variable, and one for the con
ditioning variable. We show that under mild conditions, the features are unique
up to scaling and permutation. Our results extend recent developments in nonline
ar ICA, and in fact, they lead to an important generalization of ICA models. In
particular, we show that our model can be used for the estimation of the compone
nts in the framework of Independently Modulated Component Analysis (IMCA), a new
 generalization of nonlinear ICA that relaxes the independence assumption. A tho
rough empirical study show that representations learnt by our model from real-wo
rld image datasets are identifiable, and improve performance in transfer learnin
g and semi-supervised learning tasks.
**********

Testing Determinantal Point Processes
Khashayar Gatmiry, Maryam Aliakbarpour, Stefanie Jegelka
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

CogLTX: Applying BERT to Long Texts
Ming Ding, Chang Zhou, Hongxia Yang, Jie Tang
BERTs are incapable of processing long texts due to its quadratically increasing
 memory and time consumption. The straightforward thoughts to address this probl
em, such as slicing the text by a sliding window or simplifying transformers, su
ffer from insufficient long-range attentions or need customized CUDA kernels. Th
e limited text length of BERT reminds us the limited capacity (5~ 9 chunks) of t
he working memory of humans – then how do human beings Cognize Long TeXts? Found

ed on the cognitive theory stemming from Baddeley, our CogLTX framework identifies key sentences by training a judge model, concatenates them for reasoning and enables multi-step reasoning via rehearsal and decay. Since relevance annotations are usually unavailable, we propose to use treatment experiments to create supervision. As a general algorithm, CogLTX outperforms or gets comparable results to SOTA models on NewsQA, HotpotQA, multi-class and multi-label long-text classification tasks with memory overheads independent of the text length.
**********

f-GAIL: Learning f-Divergence for Generative Adversarial Imitation Learning
Xin Zhang, Yanhua Li, Ziming Zhang, Zhi-Li Zhang
Imitation learning (IL) aims to learn a policy from expert demonstrations that minimizes the discrepancy between the learner and expert behaviors. Various imitation learning algorithms have been proposed with different pre-determined divergences to quantify the discrepancy. This naturally gives rise to the following question: Given a set of expert demonstrations, which divergence can recover the expert policy more accurately with higher data efficiency? In this work, we propose f-GAIL – a new generative adversarial imitation learning model – that automatically learns a discrepancy measure from the f-divergence family as well as a policy capable of producing expert-like behaviors. Compared with IL baselines with various predefined divergence measures, f-GAIL learns better policies with higher data efficiency in six physics-based control tasks.
**********

Non-parametric Models for Non-negative Functions
Ulysse Marteau-Ferey, Francis Bach, Alessandro Rudi
Linear models have shown great effectiveness and flexibility in many fields such as machine learning, signal processing and statistics. They can represent rich spaces of functions while preserving the convexity of the optimization problems where they are used, and are simple to evaluate, differentiate and integrate. However, for modeling non-negative functions, which are crucial for unsupervised learning, density estimation, or non-parametric Bayesian methods, linear models are not applicable directly. Moreover, current state-of-the-art models like generalized linear models either lead to non-convex optimization problems, or cannot be easily integrated. In this paper we provide the first model for non-negative functions which benefits from the same good properties of linear models. In particular, we prove that it admits a representer theorem and provide an efficient dual formulation for convex problems. We study its representation power, showing that the resulting space of functions is strictly richer than that of generalized linear models. Finally we extend the model and the theoretical results to functions with outputs in convex cones. The paper is complemented by an experimental evaluation of the model showing its effectiveness in terms of formulation, algorithmic derivation and practical results on the problems of density estimation, regression with heteroscedastic errors, and multiple quantile regression.
**********

Uncertainty Aware Semi-Supervised Learning on Graph Data
Xujiang Zhao, Feng Chen, Shu Hu, Jin-Hee Cho
Thanks to graph neural networks (GNNs), semi-supervised node classification has shown the state-of-the-art performance in graph data. However, GNNs have not considered different types of uncertainties associated with class probabilities to minimize risk of increasing misclassification under uncertainty in real life. In this work, we propose a multi-source uncertainty framework using a GNN that reflects various types of predictive uncertainties in both deep learning and belief/evidence theory domains for node classification predictions. By collecting evidence from the given labels of training nodes, the Graph-based Kernel Dirichlet distribution Estimation (GKDE) method is designed for accurately predicting node-level Dirichlet distributions and detecting out-of-distribution (OOD) nodes. We validated the outperformance of our proposed model compared to the state-of-the-art counterparts in terms of misclassification detection and OOD detection based on six real network datasets. We found that dissonance-based detection yielded the best results on misclassification detection while vacuity-based detection was the best for OOD detection. To clarify the reasons behind the results, we pr

ovided the theoretical proof that explains the relationships between different t
ypes of uncertainties considered in this work.
**********

ConvBERT: Improving BERT with Span-based Dynamic Convolution
Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, Shuicheng Yan
Pre-trained language models like BERT and its variants have recently achieved im
pressive performance in various natural language understanding tasks. However, B
ERT heavily relies on the global self-attention block and thus suffers large mem
ory footprint and computation cost. Although all its attention heads query on th
e whole input sequence for generating the attention map from a global perspectiv
e, we observe some heads only need to learn local dependencies, which means exis
tence of computation redundancy. We therefore propose a novel span-based dynamic
 convolution to replace these self-attention heads to directly model local depen
dencies. The novel convolution heads, together with the rest self-attention head
s, form a new mixed attention block that is more efficient at both global and lo
cal context learning. We equip BERT with this mixed attention design and build a
 ConvBERT model. Experiments have shown that ConvBERT significantly outperforms
BERT and its variants in various downstream tasks, with lower training cost and
fewer model parameters. Remarkably, ConvBERTbase model achieves 86.4 GLUE score,
 0.7 higher than ELECTRAbase, using less than 1/4 training cost. Code and pre-tr
ained models will be released.
**********

Practical No-box Adversarial Attacks against DNNs
Qizhang Li, Yiwen Guo, Hao Chen
The study of adversarial vulnerabilities of deep neural networks (DNNs) has prog
ressed rapidly. Existing attacks require either internal access (to the architec
ture, parameters, or training set of the victim model) or external access (to qu
ery the model). However, both the access may be infeasible or expensive in many
scenarios. We investigate no-box adversarial examples, where the attacker can ne
ither access the model information or the training set nor query the model. Inst
ead, the attacker can only gather a small number of examples from the same probl
em domain as that of the victim model. Such a stronger threat model greatly expa
nds the applicability of adversarial attacks. We propose three mechanisms for tr
aining with a very small dataset (on the order of tens of examples) and find tha
t prototypical reconstruction is the most effective. Our experiments show that a
dversarial examples crafted on prototypical auto-encoding models transfer well t
o a variety of image classification and face verification models. On a commercia
l celebrity recognition system held by clarifai.com, our approach significantly
diminishes the average prediction accuracy of the system to only 15.40%, which i
s on par with the attack that transfers adversarial examples from a pre-trained
Arcface model. Our code is publicly available at: https://github.com/qizhangli/n
obox-attacks.
**********

Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Ge
nerative Model
Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, 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.
**********

Walking in the Shadow: A New Perspective on Descent Directions for Constrained M
inimization
Hassan Mortagy, Swati Gupta, Sebastian Pokutta
Descent directions such as movement towards Frank-Wolfe vertices, away steps, in
-face away steps and pairwise directions have been an important design considera
tion in conditional gradient descent (CGD) variants. In this work, we attempt to
 demystify the impact of movement in these directions towards attaining constrai
ned minimizers. The best local direction of descent is the directional derivativ
e of the projection of the gradient, which we refer to as the "shadow" of the gr

adient. We show that the continuous-time dynamics of moving in the shadow are eq
uivalent to those of PGD however non-trivial to discretize. By projecting gradie
nts in PGD, one not only ensures feasibility but also is able to "wrap" around t
he convex region. We show that Frank-Wolfe (FW) vertices in fact recover the max
imal wrap one can obtain by projecting gradients, thus providing a new perspecti
ve to these steps. We also claim that the shadow steps give the best direction o
f descent emanating from the convex hull of all possible away-vertices. Opening
up the PGD movements in terms of shadow steps gives linear convergence, dependen
t on the number of faces. We combine these insights into a novel Shadow-CG metho
d that uses FW steps (i.e., wrap around the polytope) and shadow steps (i.e., op
timal local descent direction), while enjoying linear convergence. Our analysis
develops properties of directional derivatives of projections (which may be of i
ndependent interest), while providing a unifying view of various descent directi
ons in the CGD literature.
**********

Path Sample-Analytic Gradient Estimators for Stochastic Binary Networks
Alexander Shekhovtsov, Viktor Yanush, Boris Flach
In neural networks with binary activations and or binary weights the training by
 gradient descent is complicated as the model has piecewise constant response.
We consider stochastic binary networks, obtained by adding noises in front of ac
tivations.
The expected model response becomes a smooth function of parameters, its gradien
t is well defined but it is challenging to estimate it accurately.
We propose a new method for this estimation problem combining sampling and analy
tic approximation steps. The method has a significantly reduced variance at the
price of a small bias which gives a very practical tradeoff in comparison with e
xisting unbiased and biased estimators.
We further show that one extra linearization step leads to a deep straight-throu
gh estimator previously known only as an ad-hoc heuristic.
We experimentally show higher accuracy in gradient estimation and demonstrate a
more stable and better performing training in deep convolutional models with bot
h proposed methods.
**********

Reward Propagation Using Graph Convolutional Networks
Martin Klissarov, Doina Precup
Potential-based reward shaping provides an approach for designing good reward fu
nctions, with the purpose of speeding up learning. However, automatically findin
g potential functions for complex environments is a difficult problem (in fact,
of the same difficulty as learning a value function from scratch). We propose a
new framework for learning potential functions by leveraging ideas from graph re
presentation learning. Our approach relies on Graph Convolutional Networks which
 we use as a key ingredient in combination with the probabilistic inference view
 of reinforcement learning. More precisely, we leverage Graph Convolutional Netw
orks to perform message passing from rewarding states. The propagated messages c
an then be used as potential functions for reward shaping to accelerate learning
. We verify empirically that our approach can achieve considerable improvements
in both small and high-dimensional control problems.
**********

LoopReg: Self-supervised Learning of Implicit Surface Correspondences, Pose and
Shape for 3D Human Mesh Registration
Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, Gerard Pons-Mol
l
We address the problem of fitting 3D human models to 3D scans of dressed humans.
 Classical methods optimize both the data-to-model correspondences and the human
 model parameters (pose and shape), but are reliable only when initialised close
 to the solution. Some methods initialize the optimization based on fully superv
ised correspondence predictors, which is not differentiable end-to-end, and can
only process a single scan at a time.  Our main contribution is LoopReg, an end-
to-end learning framework to register a corpus of scans to a common 3D human mod
el.  The key idea is to create a self-supervised loop.  A backward map, paramete

rized by a Neural Network, predicts the correspondence from every scan point to the surface of the human model. A forward map, parameterized by a human model, transforms the corresponding points back to the scan based on the model parameters (pose and shape), thus closing the loop. Formulating this closed loop is not straightforward because it is not trivial to force the output of the NN to be on the surface of the human model -- outside this surface the human model is not even defined. To this end, we propose two key innovations. First, we define the canonical surface implicitly as the zero level set of a distance field in R3, which in contrast to more common UV parameterizations does not require cutting the surface, does not have discontinuities, and does not induce distortion. Second, we diffuse the human model to the 3D domain. This allows to map the NN predictions forward, even when they slightly deviate from the zero level set. Results demonstrate that we can train LoopReg mainly self-supervised -- following a supervised warm-start, the model becomes increasingly more accurate as additional unlabelled raw scans are processed. Our code and pre-trained models can be downloaded for research.
**********

Fully Dynamic Algorithm for Constrained Submodular Optimization
Silvio Lattanzi, Slobodan Mitrovi■, Ashkan Norouzi-Fard, Jakub M. Tarnawski, Morteza Zadimoghaddam
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Optimal Transport with Applications in Generative Modeling and Domain Adaptation
Yogesh Balaji, Rama Chellappa, Soheil Feizi
Optimal Transport (OT) distances such as Wasserstein have been used in several areas such as GANs and domain adaptation. OT, however, is very sensitive to outliers (samples with large noise) in the data since in its objective function, every sample, including outliers, is weighed similarly due to the marginal constraints. To remedy this issue, robust formulations of OT with unbalanced marginal constraints have previously been proposed. However, employing these methods in deep learning problems such as GANs and domain adaptation is challenging due to the instability of their dual optimization solvers. In this paper, we resolve these issues by deriving a computationally-efficient dual form of the robust OT optimization that is amenable to modern deep learning applications. We demonstrate the effectiveness of our formulation in two applications of GANs and domain adaptation. Our approach can train state-of-the-art GAN models on noisy datasets corrupted with outlier distributions. In particular, the proposed optimization method computes weights for training samples reflecting how difficult it is for those samples to be generated in the model. In domain adaptation, our robust OT formulation leads to improved accuracy compared to the standard adversarial adaptation methods. Our code is available at https://github.com/yogeshbalaji/robustOT.
**********

Autofocused oracles for model-based design
Clara Fannjiang, Jennifer Listgarten
Data-driven design is making headway into a number of application areas, including protein, small-molecule, and materials engineering. The design goal is to construct an object with desired properties, such as a protein that binds to a therapeutic target, or a superconducting material with a higher critical temperature than previously observed. To that end, costly experimental measurements are being replaced with calls to high-capacity regression models trained on labeled data, which can be leveraged in an in silico search for design candidates. However, the design goal necessitates moving into regions of the design space beyond where such models were trained. Therefore, one can ask: should the regression model be altered as the design algorithm explores the design space, in the absence of new data? Herein, we answer this question in the affirmative. In particular, we (i) formalize the data-driven design problem as a non-zero-sum game, (ii) devel

op a principled strategy for retraining the regression model as the design algor
ithm proceeds---what we refer to as autofocusing, and (iii) demonstrate the prom
ise of autofocusing empirically.
**********

Debiasing Averaged Stochastic Gradient Descent to handle missing values
Aude Sportisse, Claire Boyer, Aymeric Dieuleveut, Julie Josse
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Trajectory-wise Multiple Choice Learning for Dynamics Generalization in Reinforc
ement Learning
Younggyo Seo, Kimin Lee, Ignasi Clavera Gilaberte, Thanard Kurutach, Jinwoo Shin
, Pieter Abbeel
Model-based reinforcement learning (RL) has shown great potential in various con
trol tasks in terms of both sample-efficiency and final performance. However, le
arning a generalizable dynamics model robust to changes in dynamics remains a ch
allenge since the target transition dynamics follow a multi-modal distribution.
In this paper, we present a new model-based RL algorithm, coined trajectory-wise
 multiple choice learning, that learns a multi-headed dynamics model for dynamic
s generalization. The main idea is updating the most accurate prediction head to
 specialize each head in certain environments with similar dynamics, i.e., clust
ering environments. Moreover, we incorporate context learning, which encodes dyn
amics-specific information from past experiences into the context latent vector,
 enabling the model to perform online adaptation to unseen environments. Finally
, to utilize the specialized prediction heads more effectively, we propose an ad
aptive planning method, which selects the most accurate prediction head over a r
ecent experience. Our method exhibits superior zero-shot generalization performa
nce across a variety of control tasks, compared to state-of-the-art RL methods.
Source code and videos are available at https://sites.google.com/view/trajectory
-mcl.
**********

CompRess: Self-Supervised Learning by Compressing Representations
Soroush Abbasi Koohpayegani, Ajinkya Tejankar, Hamed Pirsiavash
Self-supervised learning aims to learn good representations with unlabeled data.
 Recent works have shown that larger models benefit more from self-supervised le
arning than smaller models. As a result, the gap between supervised and self-sup
ervised learning has been greatly reduced for larger models. In this work, inste
ad of designing a new pseudo task for self-supervised learning, we develop a mod
el compression method to compress an already learned, deep self-supervised model
 (teacher) to a smaller one (student). We train the student model so that it mim
ics the relative similarity between the datapoints in the teacher's embedding sp
ace. For AlexNet, our method outperforms all previous methods including the full
y supervised model on ImageNet linear evaluation (59.0% compared to 56.5%) and o
n nearest neighbor evaluation (50.7% compared to 41.4%). To the best of our know
ledge, this is the first time a self-supervised AlexNet has outperformed supervi
sed one on ImageNet classification. Our code is available here: https://github.c
om/UMBCvision/CompRess
**********

Sample complexity and effective dimension for regression on manifolds
Andrew McRae, Justin Romberg, Mark Davenport
We consider the theory of regression on a manifold using reproducing kernel Hilb
ert space methods. Manifold models arise in a wide variety of modern machine lea
rning problems, and our goal is to help understand the effectiveness of various
implicit and explicit dimensionality-reduction methods that exploit manifold str
ucture. Our first key contribution is to establish a novel nonasymptotic version
 of the Weyl law from differential geometry. From this we are able to show that
certain spaces of smooth functions on a manifold are effectively finite-dimensio
nal, with a complexity that scales according to the manifold dimension rather th

an any ambient data dimension. Finally, we show that given (potentially noisy) f
unction values taken uniformly at random over a manifold, a kernel regression es
timator (derived from the spectral decomposition of the manifold) yields minimax
-optimal error bounds that are controlled by the effective dimension.
**********
The phase diagram of approximation rates for deep neural networks
Dmitry Yarotsky, Anton Zhevnerchuk
We explore the phase diagram of approximation rates for deep neural networks and
 prove several new theoretical results. In particular, we generalize the existin
g result on the existence of deep discontinuous phase in ReLU networks to functi
onal classes of arbitrary positive smoothness, and identify the boundary between
 the feasible and infeasible rates. Moreover, we show that all networks with a p
iecewise polynomial activation function have the same phase diagram. Next, we de
monstrate that standard fully-connected architectures with a fixed width indepen
dent of smoothness can adapt to smoothness and achieve almost optimal rates. Fin
ally, we consider deep networks with periodic activations ("deep Fourier expansi
on") and prove that they have very fast, nearly exponential approximation rates,
 thanks to the emerging capability of the network to implement efficient lookup
operations.
**********
Timeseries Anomaly Detection using Temporal Hierarchical One-Class Network
Lifeng Shen, Zhuocong Li, James Kwok
Real-world timeseries have complex underlying temporal dynamics and the detectio
n of anomalies is challenging. In this paper, we propose the Temporal Hierarchic
al One-Class (THOC) network, a temporal one-class classification model for times
eries anomaly detection. It captures temporal dynamics in multiple scales by usi
ng a dilated recurrent neural network with skip connections. Using multiple hype
rspheres obtained with a hierarchical clustering process, a one-class objective
called Multiscale Vector Data Description is defined. This allows the temporal d
ynamics to be well captured by a set of multi-resolution temporal clusters. To f
urther facilitate representation learning, the hypersphere centers are encourage
d to be orthogonal to each other, and a self-supervision task in the temporal do
main is added. The whole model can be trained end-to-end. Extensive empirical st
udies on various real-world timeseries demonstrate that the proposed THOC networ
k outperforms recent strong deep learning baselines on timeseries anomaly detect
ion.
**********
EcoLight: Intersection Control in Developing Regions Under Extreme Budget and Ne
twork Constraints
Sachin Chauhan, Kashish Bansal, Rijurekha Sen
Effective intersection control can play an important role in reducing traffic co
ngestion and associated vehicular emissions. This is vitally needed in developin
g countries, where air pollution is reaching life threatening levels. This paper
 presents EcoLight intersection control for developing regions, where budget is
constrained and network connectivity is very poor. EcoLight learns effective con
trol offline using state-of-the-art Deep Reinforcement Learning methods, but dep
loys highly efficient runtime control algorithms on low cost embedded devices th
at work stand-alone on road without server connectivity. EcoLight optimizes both
 average case and worst case values of throughput, travel time and other metrics
, as evaluated on open-source datasets from New York and on a custom developing
region dataset.
**********
Reconstructing Perceptive Images from Brain Activity by Shape-Semantic GAN
Tao Fang, Yu Qi, Gang Pan
Reconstructing seeing images from fMRI recordings is an absorbing research area
in neuroscience and provides a potential brain-reading technology. The challenge
 lies in that visual encoding in brain is highly complex and not fully revealed.
Inspired by the theory that visual features are hierarchically represented in co
rtex, we propose to break the complex visual signals into multi-level components
 and decode each component separately. Specifically, we decode shape and semanti

c representations from the lower and higher visual cortex respectively, and merg e the shape and semantic information to images by a generative adversarial netwo rk (Shape-Semantic GAN). This 'divide and conquer' strategy captures visual info rmation more accurately. Experiments demonstrate that Shape-Semantic GAN improve s the reconstruction similarity and image quality, and achieves the state-of-the -art image reconstruction performance.
**********

Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design
Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell , Andrew Critch, Sergey Levine
A wide range of reinforcement learning (RL) problems --- including robustness, t ransfer learning, unsupervised RL, and emergent complexity --- require specifyin g a distribution of tasks or environments in which a policy will be trained. Ho wever, creating a useful distribution of environments is error prone, and takes a significant amount of developer time and effort. We propose Unsupervised Envir onment Design (UED) as an alternative paradigm, where developers provide environ ments with unknown parameters, and these parameters are used to automatically pr oduce a distribution over valid, solvable environments. Existing approaches to a utomatically generating environments suffer from common failure modes: domain ra ndomization cannot generate structure or adapt the difficulty of the environment to the agent's learning progress, and minimax adversarial training leads to wor st-case environments that are often unsolvable. To generate structured, solvable environments for our protagonist agent, we introduce a second, antagonist agent that is allied with the environment-generating adversary. The adversary is moti vated to generate environments which maximize regret, defined as the difference between the protagonist and antagonist agent's return. We call our technique Pro tagonist Antagonist Induced Regret Environment Design (PAIRED). Our experiments demonstrate that PAIRED produces a natural curriculum of increasingly complex e nvironments, and PAIRED agents achieve higher zero-shot transfer performance whe n tested in highly novel environments.
**********

A Spectral Energy Distance for Parallel Speech Synthesis
Alexey Gritsenko, Tim Salimans, Rianne van den Berg, Jasper Snoek, Nal Kalchbren ner
Speech synthesis is an important practical generative modeling problem that has seen great progress over the last few years, with likelihood-based autoregressiv e neural models now outperforming traditional concatenative systems. A downside of such autoregressive models is that they require executing tens of thousands o f sequential operations per second of generated audio, making them ill-suited fo r deployment on specialized deep learning hardware. Here, we propose a new learn ing method that allows us to train highly parallel models of speech, without req uiring access to an analytical likelihood function. Our approach is based on a g eneralized energy distance between the distributions of the generated and real a udio. This spectral energy distance is a proper scoring rule with respect to the distribution over magnitude-spectrograms of the generated waveform audio and of fers statistical consistency guarantees. The distance can be calculated from min ibatches without bias, and does not involve adversarial learning, yielding a sta ble and consistent method for training implicit generative models. Empirically, we achieve state-of-the-art generation quality among implicit generative models, as judged by the recently-proposed cFDSD metric. When combining our method with adversarial techniques, we also improve upon the recently-proposed GAN-TTS mode l in terms of Mean Opinion Score as judged by trained human evaluators.
**********

Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to I mage Perturbations
Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, James J. DiCarlo
Current state-of-the-art object recognition models are largely based on convolut ional neural network (CNN) architectures, which are loosely inspired by the prim ate visual system. However, these CNNs can be fooled by imperceptibly small, exp

licitly crafted perturbations, and struggle to recognize objects in corrupted im
ages that are easily recognized by humans. Here, by making comparisons with prim
ate neural data, we first observed that CNN models with a neural hidden layer th
at better matches primate primary visual cortex (V1) are also more robust to adv
ersarial attacks. Inspired by this observation, we developed VOneNets, a new cla
ss of hybrid CNN vision models. Each VOneNet contains a fixed weight neural netw
ork front-end that simulates primate V1, called the VOneBlock, followed by a neu
ral network back-end adapted from current CNN vision models. The VOneBlock is ba
sed on a classical neuroscientific model of V1: the linear-nonlinear-Poisson mod
el, consisting of a biologically-constrained Gabor filter bank, simple and compl
ex cell nonlinearities, and a V1 neuronal stochasticity generator. After trainin
g, VOneNets retain high ImageNet performance, but each is substantially more rob
ust, outperforming the base CNNs and state-of-the-art methods by 18% and 3%, res
pectively, on a conglomerate benchmark of perturbations comprised of white box a
dversarial attacks and common image corruptions. Finally, we show that all compo
nents of the VOneBlock work in synergy to improve robustness. While current CNN
architectures are arguably brain-inspired, the results presented here demonstrat
e that more precisely mimicking just one stage of the primate visual system lead
s to new gains in ImageNet-level computer vision applications.
**********

Learning from Positive and Unlabeled Data with Arbitrary Positive Shift
Zayd Hammoudeh, Daniel Lowd
Positive-unlabeled (PU) learning trains a binary classifier using only positive
and unlabeled data. A common simplifying assumption is that the positive data is
representative of the target positive class. This assumption rarely holds in pr
actice due to temporal drift, domain shift, and/or adversarial manipulation. Th
is paper shows that PU learning is possible even with arbitrarily non-representa
tive positive data given unlabeled data from the source and target distributions
. Our key insight is that only the negative class's distribution need be fixed.
We integrate this into two statistically consistent methods to address arbitrar
y positive bias - one approach combines negative-unlabeled learning with unlabel
ed-unlabeled learning while the other uses a novel, recursive risk estimator. Ex
perimental results demonstrate our methods' effectiveness across numerous real-w
orld datasets and forms of positive bias, including disjoint positive class-cond
itional supports. Additionally, we propose a general, simplified approach to add
ress PU risk estimation overfitting.
**********

Deep Energy-based Modeling of Discrete-Time Physics
Takashi Matsubara, Ai Ishikawa, Takaharu Yaguchi
Physical phenomena in the real world are often described by energy-based modelin
g theories, such as Hamiltonian mechanics or the Landau theory, which yield vari
ous physical laws. Recent developments in neural networks have enabled the mimic
king of the energy conservation law by learning the underlying continuous-time d
ifferential equations. However, this may not be possible in discrete time, which
is often the case in practical learning and computation. Moreover, other physic
al laws have been overlooked in the previous neural network models. In this stud
y, we propose a deep energy-based physical model that admits a specific differen
tial geometric structure. From this structure, the conservation or dissipation l
aw of energy and the mass conservation law follow naturally. To ensure the energ
etic behavior in discrete time, we also propose an automatic discrete differenti
ation algorithm that enables neural networks to employ the discrete gradient met
hod.
**********

Quantifying Learnability and Describability of Visual Concepts Emerging in Repre
sentation Learning
Iro Laina, Ruth Fong, Andrea Vedaldi
The increasing impact of black box models, and particularly of unsupervised ones
, comes with an increasing interest in tools to understand and interpret them. I
n this paper, we consider in particular how to characterise visual groupings dis
covered automatically by deep neural networks, starting with state-of-the-art cl

ustering methods. In some cases, clusters readily correspond to an existing labe
lled dataset. However, often they do not, yet they still maintain an "intuitive
interpretability''. We introduce two concepts, visual learnability and describab
ility, that can be used to quantify the interpretability of arbitrary image grou
pings, including unsupervised ones. The idea is to measure (1) how well humans c
an learn to reproduce a grouping by measuring their ability to generalise from a
 small set of visual examples (learnability) and (2) whether the set of visual e
xamples can be replaced by a succinct, textual description (describability). By
assessing human annotators as classifiers, we remove the subjective quality of e
xisting evaluation metrics. For better scalability, we finally propose a class-l
evel captioning system to generate descriptions for visual groupings automatical
ly and compare it to human annotators using the describability metric.
**********

Self-Learning Transformations for Improving Gaze and Head Redirection
Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, Otmar Hilliges
Many computer vision tasks rely on labeled data. Rapid progress in generative mo
deling has led to the ability to synthesize photorealistic images. However, cont
rolling specific aspects of the generation process such that the data can be use
d for supervision of downstream tasks remains challenging. In this paper we prop
ose a novel generative model for images of faces, that is capable of producing h
igh-quality images under fine-grained control over eye gaze and head orientation
 angles. This requires the disentangling of many appearance related factors incl
uding gaze and head orientation but also lighting, hue etc. We propose a novel a
rchitecture which learns to discover, disentangle and encode these extraneous va
riations in a self-learned manner. We further show that explicitly disentangling
 task-irrelevant factors results in more accurate modelling of gaze and head ori
entation. A novel evaluation scheme shows that our method improves upon the stat
e-of-the-art in redirection accuracy and disentanglement between gaze direction
and head orientation changes. Furthermore, we show that in the presence of limit
ed amounts of real-world training data, our method allows for improvements in th
e downstream task of semi-supervised cross-dataset gaze estimation. Please check
 our project page at: https://ait.ethz.ch/projects/2020/STED-gaze/
**********

Language-Conditioned Imitation Learning for Robot Manipulation Tasks
Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, H
eni Ben Amor
Imitation learning is a popular approach for teaching motor skills to robots. Ho
wever, most approaches focus on extracting policy parameters from execution trac
es alone (i.e., motion trajectories and perceptual data). No adequate communicat
ion channel exists between the human expert and the robot to describe critical a
spects of the task, such as the properties of the target object or the intended
shape of the motion. Motivated by insights into the human teaching process, we i
ntroduce a method for incorporating unstructured natural language into imitation
 learning. At training time, the expert can provide demonstrations along with ve
rbal descriptions in order to describe the underlying intent (e.g., "go to the l
arge green bowl"). The training process then interrelates these two modalities t
o encode the correlations between language, perception, and motion. The resultin
g language-conditioned visuomotor policies can be conditioned at runtime on new
human commands and instructions, which allows for more fine-grained control over
 the trained policies while also reducing situational ambiguity. We demonstrate
in a set of simulation experiments how our approach can learn language-condition
ed manipulation policies for a seven-degree-of-freedom robot arm and compare the
 results to a variety of alternative methods.
**********

POMDPs in Continuous Time and Discrete Spaces
Bastian Alt, Matthias Schultheis, Heinz Koeppl
Many processes, such as discrete event systems in engineering or population dyna
mics in biology, evolve in discrete space and continuous time. We consider the p
roblem of optimal decision making in such discrete state and action space system
s under partial observability. This places our work at the intersection of optim

al filtering and optimal control. At the current state of research, a mathematic al description for simultaneous decision making and filtering in continuous time with finite state and action spaces is still missing. In this paper, we give a mathematical description of a continuous-time partial observable Markov decision process (POMDP). By leveraging optimal filtering theory we derive a Hamilton-Ja cobi-Bellman (HJB) type equation that characterizes the optimal solution. Using techniques from deep learning we approximately solve the resulting partial integ ro-differential equation. We present (i) an approach solving the decision proble m offline by learning an approximation of the value function and (ii) an online algorithm which provides a solution in belief space using deep reinforcement lea rning. We show the applicability on a set of toy examples which pave the way for future methods providing solutions for high dimensional problems.
**********

Exemplar Guided Active Learning

Jason S. Hartford, Kevin Leyton-Brown, Hadas Raviv, Dan Padnos, Shahar Lev, Bara k Lenz

We consider the problem of wisely using a limited budget to label a small subset of a large unlabeled dataset. For example, consider the NLP problem of word sen se disambiguation. For any word, we have a set of candidate labels from a knowle dge base, but the label set is not necessarily representative of what occurs in the data: there may exist labels in the knowledge base that very rarely occur in the corpus because the sense is rare in modern English; and conversely there ma y exist true labels that do not exist in our knowledge base. Our aim is to obtai n a classifier that performs as well as possible on examples of each “common cla ss” that occurs with frequency above a given threshold in the unlabeled set whil e annotating as few examples as possible from “rare classes” whose labels occur with less than this frequency. The challenge is that we are not informed which l abels are common and which are rare, and the true label distribution may exhibit extreme skew. We describe an active learning approach that (1) explicitly searc hes for rare classes by leveraging the contextual embedding spaces provided by m odern language models, and (2) incorporates a stopping rule that ignores classes once we prove that they occur below our target threshold with high probability. We prove that our algorithm only costs logarithmically more than a hypothetical approach that knows all true label frequencies and show experimentally that inc orporating automated search can significantly reduce the number of samples neede d to reach target accuracy levels.
**********

Grasp Proposal Networks: An End-to-End Solution for Visual Learning of Robotic G rasps

Chaozheng Wu, Jian Chen, Qiaoyu Cao, Jianchi Zhang, Yunxin Tai, Lin Sun, Kui Jia

Learning robotic grasps from visual observations is a promising yet challenging task. Recent research shows its great potential by preparing and learning from l arge-scale synthetic datasets. For the popular, 6 degree-of-freedom (6-DOF) gras p setting of parallel-jaw gripper, most of existing methods take the strategy of heuristically sampling grasp candidates and then evaluating them using learned scoring functions. This strategy is limited in terms of the conflict between sam pling efficiency and coverage of optimal grasps. To this end, we propose in this work a novel, end-to-end \emph{Grasp Proposal Network (GPNet)}, to predict a di verse set of 6-DOF grasps for an unseen object observed from a single and unknow n camera view. GPNet builds on a key design of grasp proposal module that define s \emph{anchors of grasp centers} at discrete but regular 3D grid corners, which is flexible to support either more precise or more diverse grasp predictions. T o test GPNet, we contribute a synthetic dataset of 6-DOF object grasps; evaluati on is conducted using rule-based criteria, simulation test, and real test. Compa rative results show the advantage of our methods over existing ones. Notably, GP Net gains better simulation results via the specified coverage, which helps achi eve a ready translation in real test. Our code and dataset are available on \url {https://github.com/CZ-Wu/GPNet}.
**********

Node Embeddings and Exact Low-Rank Representations of Complex Networks

Sudhanshu Chanpuriya, Cameron Musco, Konstantinos Sotiropoulos, Charalampos Tsourakakis

Low-dimensional embeddings, from classical spectral embeddings to modern neural-net-inspired methods, are a cornerstone in the modeling and analysis of complex networks. Recent work by Seshadhri et al. (PNAS 2020) suggests that such embeddings cannot capture local structure arising in complex networks. In particular, they show that any network generated from a natural low-dimensional model cannot be both sparse and have high triangle density (high clustering coefficient), two hallmark properties of many real-world networks.

**********

Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications

Sarah Perrin, Julien Perolat, Mathieu Lauriere, Matthieu Geist, Romuald Elie, Olivier Pietquin

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

**********

Steering Distortions to Preserve Classes and Neighbors in Supervised Dimensionality Reduction

Benoît Colange, Jaakko Peltonen, Michael Aupetit, Denys Dutykh, Sylvain Lespinats

Nonlinear dimensionality reduction of high-dimensional data is challenging as the low-dimensional embedding will necessarily contain distortions, and it can be hard to determine which distortions are the most important to avoid. When annotation of data into known relevant classes is available, it can be used to guide the embedding to avoid distortions that worsen class separation. The supervised mapping method introduced in the present paper, called ClassNeRV, proposes an original stress function that takes class annotation into account and evaluates embedding quality both in terms of false neighbors and missed neighbors. ClassNeRV shares the theoretical framework of a family of methods descended from Stochastic Neighbor Embedding (SNE). Our approach has a key advantage over previous ones: in the literature supervised methods often emphasize class separation at the price of distorting the data neighbors' structure; conversely, unsupervised methods provide better preservation of structure at the price of often mixing classes. Experiments show that ClassNeRV can preserve both neighbor structure and class separation, outperforming nine state of the art alternatives.

**********

On Infinite-Width Hypernetworks

Etai Littwin, Tomer Galanti, Lior Wolf, Greg Yang

{\em Hypernetworks} are architectures that produce the weights of a task-specific {\em primary network}. A notable application of hypernetworks in the recent literature involves learning to output functional representations. In these scenarios, the hypernetwork learns a representation corresponding to the weights of a shallow MLP, which typically encodes shape or image information. While such representations have seen considerable success in practice, they remain lacking in the theoretical guarantees in the wide regime of the standard architectures. In this work, we study wide over-parameterized hypernetworks. We show that unlike typical architectures, infinitely wide hypernetworks do not guarantee convergence to a global minima under gradient descent. We further show that convexity can be achieved by increasing the dimensionality of the hypernetwork's output, to represent wide MLPs. In the dually infinite-width regime, we identify the functional priors of these architectures by deriving their corresponding GP and NTK kernels, the latter of which we refer to as the {\em hyperkernel}. As part of this study, we make a mathematical contribution by deriving tight bounds on high order Taylor expansion terms of standard fully connected ReLU networks.

**********

Interferobot: aligning an optical interferometer by a reinforcement learning agent

Dmitry Sorokin, Alexander Ulanov, Ekaterina Sazhina, Alexander Lvovsky

Limitations in acquiring training data restrict potential applications of deep r
einforcement learning (RL) methods to the training of real-world robots. Here we
 train an RL agent to align a Mach-Zehnder interferometer, which is an essential
 part of many optical experiments, based on images of interference fringes acqui
red by a monocular camera. The agent is trained in a simulated environment, with
out any hand-coded features or a priori information about the physics, and subse
quently transferred to a physical interferometer. Thanks to a set of domain rand
omizations simulating uncertainties in physical measurements, the agent successf
ully aligns this interferometer without any fine-tuning, achieving a performance
 level of a human expert.
**********

Program Synthesis with Pragmatic Communication
Yewen Pu, Kevin Ellis, Marta Kryven, Josh Tenenbaum, Armando Solar-Lezama
Program synthesis techniques construct or infer programs from user-provided spec
ifications, such as input-output examples. Yet most specifications, especially t
hose given by end-users, leave the synthesis problem radically ill-posed, becaus
e many programs may simultaneously satisfy the specification. Prior work resolve
s this ambiguity by using various inductive biases, such as a preference for sim
pler programs. This work introduces a new inductive bias derived by modeling the
 program synthesis task as rational communication, drawing insights from recursi
ve reasoning models of pragmatics. Given a specification, we score a candidate p
rogram both on its consistency with the specification, and also whether a ration
al speaker would chose this particular specification to communicate that program
. We develop efficient algorithms for such an approach when learning from input-
output examples, and build a pragmatic program synthesizer over a simple grid-li
ke layout domain. A user study finds that end-user participants communicate more
 effectively with the pragmatic program synthesizer over a non-pragmatic one.
**********

Principal Neighbourhood Aggregation for Graph Nets
Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, Petar Veli■kovi■
Graph Neural Networks (GNNs) have been shown to be effective models for differen
t predictive tasks on graph-structured data. Recent work on their expressive pow
er has focused on isomorphism tasks and countable feature spaces. We extend this
 theoretical framework to include continuous features---which occur regularly in
 real-world input domains and within the hidden layers of GNNs---and we demonstr
ate the requirement for multiple aggregation functions in this context. Accordin
gly, we propose Principal Neighbourhood Aggregation (PNA), a novel architecture
combining multiple aggregators with degree-scalers (which generalize the sum agg
regator). Finally, we compare the capacity of different models to capture and ex
ploit the graph structure via a novel benchmark containing multiple tasks taken
from classical graph theory, alongside existing benchmarks from real-world domai
ns, all of which demonstrate the strength of our model. With this work we hope t
o steer some of the GNN research towards new aggregation methods which we believ
e are essential in the search for powerful and robust models.
**********

Reliable Graph Neural Networks via Robust Aggregation
Simon Geisler, Daniel Zügner, Stephan Günnemann
Perturbations targeting the graph structure have proven to be extremely effectiv
e in reducing the performance of Graph Neural Networks (GNNs), and traditional d
efenses such as adversarial training do not seem to be able to improve robustnes
s. This work is motivated by the observation that adversarially injected edges e
ffectively can be viewed as additional samples to a node's neighborhood aggregat
ion function, which results in distorted aggregations accumulating over the laye
rs. Conventional GNN aggregation functions, such as a sum or mean, can be distor
ted arbitrarily by a single outlier. We propose a robust aggregation function mo
tivated by the field of robust statistics. Our approach exhibits the largest pos
sible breakdown point of 0.5, which means that the bias of the aggregation is bo
unded as long as the fraction of adversarial edges of a node is less than 50%. O
ur novel aggregation function, Soft Medoid, is a fully differentiable generaliza
tion of the Medoid and therefore lends itself well for end-to-end deep learning.

Equipping a GNN with our aggregation improves the robustness with respect to st
ructure perturbations on Cora ML by a factor of 3 (and 5.5 on Citeseer) and by a
factor of 8 for low-degree nodes.
**********

Instance Selection for GANs
Terrance DeVries, Michal Drozdzal, Graham W. Taylor
Recent advances in Generative Adversarial Networks (GANs) have led to their wide
spread adoption for the purposes of generating high quality synthetic imagery. W
hile capable of generating photo-realistic images, these models often produce un
realistic samples which fall outside of the data manifold. Several recently prop
osed techniques attempt to avoid spurious samples, either by rejecting them afte
r generation, or by truncating the model's latent space. While effective, these
methods are inefficient, as a large fraction of training time and model capacity
are dedicated towards samples that will ultimately go unused. In this work we p
ropose a novel approach to improve sample quality: altering the training dataset
via instance selection before model training has taken place. By refining the e
mpirical data distribution before training, we redirect model capacity towards h
igh-density regions, which ultimately improves sample fidelity, lowers model cap
acity requirements, and significantly reduces training time. Code is available a
t https://github.com/uoguelph-mlrg/instanceselectionfor_gans.
**********

Linear Disentangled Representations and Unsupervised Action Estimation
Matthew Painter, Adam Prugel-Bennett, Jonathon Hare
Disentangled representation learning has seen a surge in interest over recent ti
mes, generally focusing on new models which optimise one of many disparate disen
tanglement metrics. Symmetry Based Disentangled Representation learning introduc
ed a robust mathematical framework that defined precisely what is meant by a ``l
inear disentangled representation''. This framework determined that such represe
ntations would depend on a particular decomposition of the symmetry group acting
on the data, showing that actions would manifest through irreducible group repr
esentations acting on independent representational subspaces. \citet{forwardvae}
subsequently proposed the first model to induce and demonstrate a linear disent
angled representation in a VAE model. In this work we empirically show that line
ar disentangled representations are not present in standard VAE models and that
they instead require altering the loss landscape to induce them. We proceed to s
how that such representations are a desirable property with regard to classical
disentanglement metrics. Finally we propose a method to induce irreducible repre
sentations which forgoes the need for labelled action sequences, as was required
by prior work. We explore a number of properties of this method, including the
ability to learn from action sequences without knowledge of intermediate states
and robustness under visual noise. We also demonstrate that it can successfully
learn 4 independent symmetries directly from pixels.
**********

Video Frame Interpolation without Temporal Priors
Youjian Zhang, Chaoyue Wang, Dacheng Tao
Video frame interpolation, which aims to synthesize non-exist intermediate frame
s in a video sequence, is an important research topic in computer vision. Existi
ng video frame interpolation methods have achieved remarkable results under spec
ific assumptions, such as instant or known exposure time. However, in complicate
d real-world situations, the temporal priors of videos, i.e. frames per second (
FPS) and frame exposure time, may vary from different camera sensors. When test
videos are taken under different exposure settings from training ones, the inter
polated frames will suffer significant misalignment problems. In this work, we s
olve the video frame interpolation problem in a general situation, where input f
rames can be acquired under uncertain exposure (and interval) time. Unlike previ
ous methods that can only be applied to a specific temporal prior, we derive a g
eneral curvilinear motion trajectory formula from four consecutive sharp frames
or two consecutive blurry frames without temporal priors. Moreover, utilizing co
nstraints within adjacent motion trajectories, we devise a novel optical flow re
finement strategy for better interpolation results. Finally, experiments demonst

rate that one well-trained model is enough for synthesizing high-quality slow-mo tion videos under complicated real-world situations. Codes are available on http s://github.com/yjzhang96/UTI-VFI.
**********

Learning compositional functions via multiplicative weight updates
Jeremy Bernstein, Jiawei Zhao, Markus Meister, Ming-Yu Liu, Anima Anandkumar, Yi song Yue
Compositionality is a basic structural feature of both biological and artificial neural networks. Learning compositional functions via gradient descent incurs w ell known problems like vanishing and exploding gradients, making careful learni ng rate tuning essential for real-world applications. This paper proves that mul tiplicative weight updates satisfy a descent lemma tailored to compositional fun ctions. Based on this lemma, we derive Madam---a multiplicative version of the A dam optimiser---and show that it can train state of the art neural network archi tectures without learning rate tuning. We further show that Madam is easily adap ted to train natively compressed neural networks by representing their weights i n a logarithmic number system. We conclude by drawing connections between multip licative weight updates and recent findings about synapses in biology.
**********

Sample Complexity of Uniform Convergence for Multicalibration
Eliran Shabat, Lee Cohen, Yishay Mansour
There is a growing interest in societal concerns in machine learning systems, especially in fairness.
Multicalibration gives a comprehensive methodology to address group fairness.
In this work, we address the multicalibration error and decouple it from the pre diction error. The importance of decoupling the fairness metric (multicalibratio n) and the accuracy (prediction error) is due to the inherent trade-off between the two, and the societal decision regarding the ``right tradeoff'' (as imposed many times by regulators).
Our work gives sample complexity bounds for uniform convergence guarantees of mu lticalibration error, which implies that regardless of the accuracy, we can guar antee that the empirical and (true) multicalibration errors are close.
We emphasize that our results: (1) are more general than previous bounds, as the y apply to both agnostic and realizable settings, and do not rely on a specific type of algorithm (such as differentially private), (2) improve over previous mu lticalibration sample complexity bounds and (3) implies uniform convergence guar antees for the classical calibration error.
**********

Differentiable Neural Architecture Search in Equivalent Space with Exploration E nhancement
Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Zongyuan Ge, Steven Su
Recent works on One-Shot Neural Architecture Search (NAS) mostly adopt a bilevel optimization scheme to alternatively optimize the supernet weights and architec ture parameters after relaxing the discrete search space into a differentiable s pace. However, the non-negligible incongruence in their relaxation methods is ha rd to guarantee the differentiable optimization in the continuous space is equiv alent to the optimization in the discrete space. Differently, this paper utilize s a variational graph autoencoder to injectively transform the discrete architec ture space into an equivalently continuous latent space, to resolve the incongru ence. A probabilistic exploration enhancement method is accordingly devised to e ncourage intelligent exploration during the architecture search in the latent sp ace, to avoid local optimal in architecture search. As the catastrophic forgetti ng in differentiable One-Shot NAS deteriorates supernet predictive ability and m akes the bilevel optimization inefficient, this paper further proposes an archi tecture complementation method to relieve this deficiency. We analyze the effect iveness of the proposed method, and a series of experiments have been conducted to compare the proposed method with state-of-the-art One-Shot NAS methods.
**********

The interplay between randomness and structure during learning in RNNs
Friedrich Schuessler, Francesca Mastrogiuseppe, Alexis Dubreuil, Srdjan Ostojic,

Omri Barak
Training recurrent neural networks (RNNs) on low-dimensional tasks has been widely used to model functional biological networks. However, the solutions found by learning and the effect of initial connectivity are not well understood. Here, we examine RNNs trained using gradient descent on different tasks inspired by the neuroscience literature. We find that the changes in recurrent connectivity can be described by low-rank matrices. This observation holds even in the presence of random initial connectivity, although this initial connectivity has full rank and significantly accelerates training. To understand the origin of these observations, we turn to an analytically tractable setting: training a linear RNN on a simpler task. We show how the low-dimensional task structure leads to low-rank changes to connectivity, and how random initial connectivity facilitates learning. Altogether, our study opens a new perspective to understand learning in RNNs in light of low-rank connectivity changes and the synergistic role of random initialization.
**********
A Generalized Neural Tangent Kernel Analysis for Two-layer Neural Networks
Zixiang Chen, Yuan Cao, Quanquan Gu, Tong Zhang
A recent breakthrough in deep learning theory shows that the training of over-parameterized deep neural networks can be characterized by a kernel function called \textit{neural tangent kernel} (NTK). However, it is known that this type of results does not perfectly match the practice, as NTK-based analysis requires the network weights to stay very close to their initialization throughout training, and cannot handle regularizers or gradient noises. In this paper, we provide a generalized neural tangent kernel analysis and show that noisy gradient descent with weight decay can still exhibit a ``kernel-like'' behavior. This implies that the training loss converges linearly up to a certain accuracy. We also establish a novel generalization error bound for two-layer neural networks trained by noisy gradient descent with weight decay.
**********
Instance-wise Feature Grouping
Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, Jennifer Dy
In many learning problems, the domain scientist is often interested in discovering the groups of features that are redundant and are important for classification.  Moreover, the features that belong to each group, and the important feature groups may vary per sample.  But what do we mean by feature redundancy?  In this paper, we formally define two types of redundancies using information theory: \textit{Representation} and \textit{Relevant redundancies}.  We leverage these redundancies to design a formulation for instance-wise feature group discovery and reveal a theoretical guideline to help discover the appropriate number of groups.  We approximate mutual information via a variational lower bound and learn the feature group and selector indicators with Gumbel-Softmax in optimizing our formulation. Experiments on synthetic data validate our theoretical claims. Experiments on MNIST, Fashion MNIST, and gene expression datasets show that our method discovers feature groups with high classification accuracies.
**********
Robust Disentanglement of a Few Factors at a Time using rPU-VAE
Benjamin Estermann, Markus Marks, Mehmet Fatih Yanik
Disentanglement is at the forefront of unsupervised learning, as disentangled representations of data improve generalization, interpretability, and performance in downstream tasks.  Current unsupervised approaches remain inapplicable for real-world datasets since they are highly variable in their performance and fail to reach levels of disentanglement of (semi-)supervised approaches. We introduce population-based training (PBT) for improving consistency in training variational autoencoders (VAEs) and demonstrate the validity of this approach in a supervised setting (PBT-VAE). We then use Unsupervised Disentanglement Ranking (UDR) as an unsupervised heuristic to score models in our PBT-VAE training and show how models trained this way tend to consistently disentangle only a subset of the generative factors. Building on top of this observation we introduce the recursive rPU-VAE approach. We train the model until convergence, remove the lea

rned factors from the dataset and reiterate.  In doing so, we can label subsets of the dataset with the learned factors and consecutively use these labels to train one model that fully disentangles the whole dataset.  With this approach, we show striking improvement in state-of-the-art unsupervised disentanglement performance and robustness across multiple datasets and metrics.
**********
PC-PG: Policy Cover Directed Exploration for Provable Policy Gradient Learning
Alekh Agarwal, Mikael Henaff, Sham Kakade, Wen Sun
Direct policy gradient methods for reinforcement learning are a successful approach for a variety of reasons: they are model free, they directly optimize the performance metric of interest, and they allow for richly parameterized policies. Their primary drawback is that, by being local in nature, they fail to adequately explore the environment. In contrast, while model-based approaches and Q-learning can, at least in theory, directly handle exploration through the use of optimism, their ability to handle model misspecification and function approximation is far less evident. This work introduces the the POLICY COVER GUIDED POLICY GRADIENT (PC- PG) algorithm, which provably balances the exploration vs. exploitation tradeoff using an ensemble of learned policies (the policy cover). PC-PG enjoys polynomial sample complexity and run time for both tabular MDPs and, more generally, linear MDPs in an infinite dimensional RKHS. Furthermore, PC-PG also has strong guarantees under model misspecification that go beyond the standard worst case L infinity assumptions; these include approximation guarantees for state aggregation under an average case error assumption, along with guarantees under a more general assumption where the approximation error under distribution shift is controlled. We complement the theory with empirical evaluation across a variety of domains in both reward-free and reward-driven settings.
**********
Group Contextual Encoding for 3D Point Clouds
Xu Liu, Chengtao Li, Jian Wang, Jingbo Wang, Boxin Shi, Xiaodong He
Global context is crucial for 3D point cloud scene understanding tasks. In this work, we extended the contextual encoding layer that was originally designed for 2D tasks to 3D Point Cloud scenarios.  The encoding layer learns a set of code words in the feature space of the 3D point cloud to characterize the global semantic context, and then based on these code words, the method learns a global contextual descriptor to reweight the featuremaps accordingly. Moreover, compared to 2D scenarios, data sparsity becomes a major issue in 3D point cloud scenarios, and the performance of contextual encoding quickly saturates when the number of code words increases. To mitigate this problem, we further proposed a group contextual encoding method, which divides the channel into groups and then performs encoding on group-divided feature vectors. This method facilitates learning of global context in grouped subspace for 3D point clouds. We evaluate the effectiveness and generalizability of our method on three widely-studied 3D point cloud tasks. Experimental results have shown that the proposed method outperformed the VoteNet  remarkably with 3 mAP on the benchmark of SUN-RGBD, with the metrics of mAP@ 0.25, and a much greater margin of 6.57 mAP on  ScanNet  with the metrics of  mAP@ 0.5. Compared to the baseline of PointNet++,  the proposed method leads to an accuracy of 86 %, outperforming the baseline by 1.5 %. Our proposed method have outperformed the non-grouping baseline methods across the board  and establishes new state-of-the-art on these benchmarks.
**********
Latent Bandits Revisited
Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, Craig Boutilier
A latent bandit is a bandit problem where the learning agent knows reward distributions of arms conditioned on an unknown discrete latent state. The goal of the agent is to identify the latent state, after which it can act optimally. This setting is a natural midpoint between online and offline learning, where complex models can be learned offline and the agent identifies the latent state online. This is of high practical relevance, for instance in recommender systems. In this work, we propose general algorithms for latent bandits, based on both upper co

nfidence bounds and Thompson sampling. The algorithms are contextual, and aware of model uncertainty and misspecification. We provide a unified theoretical analysis of our algorithms, which have lower regret than classic bandit policies when the number of latent states is smaller than actions. A comprehensive empirical study showcases the advantages of our approach.

**********

Is normalization indispensable for training deep neural network?

Jie Shao, Kai Hu, Changhu Wang, Xiangyang Xue, Bhiksha Raj

Normalization operations are widely used to train deep neural networks, and they can improve both convergence and generalization in most tasks. The theories for normalization's effectiveness and new forms of normalization have always been hot topics in research. To better understand normalization, one question can be whether normalization is indispensable for training deep neural network? In this paper, we study what would happen when normalization layers are removed from the network, and show how to train deep neural networks without normalization layers and without performance degradation. Our proposed method can achieve the same or even slightly better performance in a variety of tasks: image classification in ImageNet, object detection and segmentation in MS-COCO, video classification in Kinetics, and machine translation in WMT English-German, etc. Our study may help better understand the role of normalization layers and can be a competitive alternative to normalization layers. Codes are available.

**********

Optimization and Generalization of Shallow Neural Networks with Quadratic Activation Functions

Stefano Sarao Mannelli, Eric Vanden-Eijnden, Lenka Zdeborová

We study the dynamics of optimization and the generalization properties of one-hidden layer neural networks with quadratic activation function in the overparametrized regime where the layer width m is larger than the input dimension d.

**********

Intra Order-preserving Functions for Calibration of Multi-Class Neural Networks

Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, Byron Boots

Predicting calibrated confidence scores for multi-class deep networks is important for avoiding rare but costly mistakes. A common approach is to learn a post-hoc calibration function that transforms the output of the original network into calibrated confidence scores while maintaining the network's accuracy. However, previous post-hoc calibration techniques work only with simple calibration functions, potentially lacking sufficient representation to calibrate the complex function landscape of deep networks. In this work, we aim to learn general post-hoc calibration functions that can preserve the top-k predictions of any deep network. We call this family of functions intra order-preserving functions. We propose a new neural network architecture that represents a class of intra order-preserving functions by combining common neural network components. Additionally, we introduce order-invariant and diagonal sub-families, which can act as regularization for better generalization when the training data size is small. We show the effectiveness of the proposed method across a wide range of datasets and classifiers. Our method outperforms state-of-the-art post-hoc calibration methods, namely temperature scaling and Dirichlet calibration, in several evaluation metrics for the task.

**********

Linear Time Sinkhorn Divergences using Positive Features

Meyer Scetbon, Marco Cuturi

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

**********

VarGrad: A Low-Variance Gradient Estimator for Variational Inference

Lorenz Richter, Ayman Boustati, Nikolas Nüsken, Francisco Ruiz, Omer Deniz Akyildiz

We analyse the properties of an unbiased gradient estimator of the ELBO for vari

ational inference, based on the score function method with leave-one-out control variates. We show that this gradient estimator can be obtained using a new loss, defined as the variance of the log-ratio between the exact posterior and the variational approximation, which we call the log-variance loss. Under certain conditions, the gradient of the log-variance loss equals the gradient of the (negative) ELBO. We show theoretically that this gradient estimator, which we call VarGrad due to its connection to the log-variance loss, exhibits lower variance than the score function method in certain settings, and that the leave-one-out control variate coefficients are close to the optimal ones. We empirically demonstrate that VarGrad offers a favourable variance versus computation trade-off compared to other state-of-the-art estimators on a discrete VAE.
**********

A Convolutional Auto-Encoder for Haplotype Assembly and Viral Quasispecies Reconstruction
Ziqi Ke, Haris Vikalo
Haplotype assembly and viral quasispecies reconstruction are challenging tasks concerned with analysis of genomic mixtures using sequencing data. High-throughput sequencing technologies generate enormous amounts of short fragments (reads) which essentially oversample components of a mixture; the representation redundancy enables reconstruction of the components (haplotypes, viral strains). The reconstruction problem, known to be NP-hard, boils down to grouping together reads originating from the same component in a mixture. Existing methods struggle to solve this problem with required level of accuracy and low runtimes; the problem is becoming increasingly more challenging as the number and length of the components increase. This paper proposes a read clustering method based on a convolutional auto-encoder designed to first project sequenced fragments to a low-dimensional space and then estimate the probability of the read origin using learned embedded features. The components are reconstructed by finding consensus sequences that agglomerate reads from the same origin. Mini-batch stochastic gradient descent and dimension reduction of reads allow the proposed method to efficiently deal with massive numbers of long reads. Experiments on simulated, semi-experimental and experimental data demonstrate the ability of the proposed method to accurately reconstruct haplotypes and viral quasispecies, often demonstrating superior performance compared to state-of-the-art methods. Source codes are available at https://github.com/WuLoli/CAECseq.
**********

Promoting Stochasticity for Expressive Policies via a Simple and Efficient Regularization Method
Qi Zhou, Yufei Kuang, Zherui Qiu, Houqiang Li, Jie Wang
Many recent reinforcement learning (RL) methods learn stochastic policies with entropy regularization for exploration and robustness. However, in continuous action spaces, integrating entropy regularization with expressive policies is challenging and usually requires complex inference procedures. To tackle this problem, we propose a novel regularization method that is compatible with a broad range of expressive policy architectures. An appealing feature is that, the estimation of our regularization terms is simple and efficient even when the policy distributions are unknown. We show that our approach can effectively promote the exploration in continuous action spaces. Based on our regularization, we propose an off-policy actor-critic algorithm. Experiments demonstrate that the proposed algorithm outperforms state-of-the-art regularized RL methods in continuous control tasks.
**********

Adversarial Counterfactual Learning and Evaluation for Recommender System
Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, Kannan Achan
The feedback data of recommender systems are often subject to what was exposed to the users; however, most learning and evaluation methods do not account for the underlying exposure mechanism. We first show in theory that applying supervised learning to detect user preferences may end up with inconsistent results in the absence of exposure information. The counterfactual propensity-weighting approach from causal inference can account for the exposure mechanism; nevertheless,

the partial-observation nature of the feedback data can cause identifiability is sues. We propose a principled solution by introducing a minimax empirical risk f ormulation. We show that the relaxation of the dual problem can be converted to an adversarial game between two recommendation models, where the opponent of the candidate model characterizes the underlying exposure mechanism. We provide lea rning bounds and conduct extensive simulation studies to illustrate and justify the proposed approach over a broad range of recommendation settings, which shed insights on the various benefits of the proposed approach.
**********
Memory-Efficient Learning of Stable Linear Dynamical Systems for Prediction and Control
Giorgos ('Yorgos') Mamakoukas, Orest Xherija, Todd Murphey
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Evolving Normalization-Activation Layers
Hanxiao Liu, Andy Brock, Karen Simonyan, Quoc Le
Normalization layers and activation functions are fundamental components in deep networks and typically co-locate with each other. Here we propose to design the m using an automated approach. Instead of designing them separately, we unify th em into a single tensor-to-tensor computation graph, and evolve its structure st arting from basic mathematical functions. Examples of such mathematical function s are addition, multiplication and statistical moments. The use of low-level mat hematical functions, in contrast to the use of high-level modules in mainstream NAS, leads to a highly sparse and large search space which can be challenging fo r search methods. To address the challenge, we develop efficient rejection proto cols to quickly filter out candidate layers that do not work well. We also use m ulti-objective evolution to optimize each layer's performance across many archit ectures to prevent overfitting. Our method leads to the discovery of EvoNorms, a set of new normalization-activation layers with novel, and sometimes surprising structures that go beyond existing design patterns. For example, some EvoNorms do not assume that normalization and activation functions must be applied sequen tially, nor need to center the feature maps, nor require explicit activation fun ctions. Our experiments show that EvoNorms work well on image classification mod els including ResNets, MobileNets and EfficientNets but also transfer well to Ma sk R-CNN with FPN/SpineNet for instance segmentation and to BigGAN for image syn thesis, outperforming BatchNorm and GroupNorm based layers in many cases.
**********
ScaleCom: Scalable Sparsified Gradient Compression for Communication-Efficient D istributed Training
Chia-Yu Chen, Jiamin Ni, Songtao Lu, Xiaodong Cui, Pin-Yu Chen, Xiao Sun, Naigan g Wang, Swagath Venkataramani, Vijayalakshmi (Viji) Srinivasan, Wei Zhang, Kaila sh Gopalakrishnan
Large-scale distributed training of Deep Neural Networks (DNNs) on state-of-the- art platforms are expected to be severely communication constrained. To overcome this limitation, numerous gradient compression techniques have been proposed an d have demonstrated high compression ratios. However, most existing compression methods do not scale well to large scale distributed systems (due to gradient bu ild-up) and / or lack evaluations in large datasets. To mitigate these issues, w e propose a new compression technique, Scalable Sparsified Gradient Compression (ScaleComp), that (i) leverages similarity in the gradient distribution amongst learners to provide a commutative compressor and keep communication cost constan t to worker number and (ii) includes low-pass filter in local gradient accumulat ions to mitigate the impacts of large batch size training and significantly impr ove scalability. Using theoretical analysis, we show that ScaleComp provides fav orable convergence guarantees and is compatible with gradient all-reduce techniq ues. Furthermore, we experimentally demonstrate that ScaleComp has small overhe ads, directly reduces gradient traffic and provides high compression rates (70-1

50X) and excellent scalability (up to 64-80 learners and 10X larger batch sizes over normal training) across a wide range of applications (image, language, and speech) without significant accuracy loss.
**********

RelationNet++: Bridging Visual Representations for Object Detection via Transformer Decoder
Cheng Chi, Fangyun Wei, Han Hu
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Learning of Discrete Graphical Models
Marc Vuffray, Sidhant Misra, Andrey Lokhov
Graphical models are useful tools for describing structured high-dimensional probability distributions. Development of efficient algorithms for learning graphical models with least amount of data remains an active research topic. Reconstruction of graphical models that describe the statistics of discrete variables is a particularly challenging problem, for which the maximum likelihood approach is intractable.
In this work, we provide the first sample-efficient method based on the Interaction Screening framework that allows one to provably learn fully general discrete factor models with node-specific discrete alphabets and multi-body interactions, specified in an arbitrary basis. We identify a single condition related to model parametrization that leads to rigorous guarantees on the recovery of model structure and parameters in any error norm, and is readily verifiable for a large class of models. Importantly, our bounds make explicit distinction between parameters that are proper to the model and priors used as an input to the algorithm.
Finally, we show that the Interaction Screening framework includes all models previously considered in the literature as special cases, and for which our analysis shows a systematic improvement in sample complexity.
**********

Near-Optimal SQ Lower Bounds for Agnostically Learning Halfspaces and ReLUs under Gaussian Marginals
Ilias Diakonikolas, Daniel Kane, Nikos Zarifis
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Neurosymbolic Transformers for Multi-Agent Communication
Jeevana Priya Inala, Yichen Yang, James Paulos, Yewen Pu, Osbert Bastani, Vijay Kumar, Martin Rinard, Armando Solar-Lezama
We study the problem of inferring communication structures that can solve cooperative multi-agent planning problems while minimizing the amount of communication. We quantify the amount of communication as the maximum degree of the communication graph; this metric captures settings where agents have limited bandwidth. Minimizing communication is challenging due to the combinatorial nature of both the decision space and the objective; for instance, we cannot solve this problem by training neural networks using gradient descent. We propose a novel algorithm that synthesizes a control policy that combines a programmatic communication policy used to generate the communication graph with a transformer policy network used to choose actions. Our algorithm first trains the transformer policy, which implicitly generates a "soft" communication graph; then, it synthesizes a programmatic communication policy that "hardens" this graph, forming a neurosymbolic transformer. Our experiments demonstrate how our approach can synthesize policies that generate low-degree communication graphs while maintaining near-optimal performance.
**********

Fairness in Streaming Submodular Maximization: Algorithms and Hardness
Marwa El Halabi, Slobodan Mitrovi■, Ashkan Norouzi-Fard, Jakab Tardos, Jakub M. Tarnawski
Submodular maximization has become established as the method of choice for the task of selecting representative and diverse summaries of data. However, if datapoints have sensitive attributes such as gender or age, such machine learning algorithms, left unchecked, are known to exhibit bias: under- or over-representation of particular groups. This has made the design of fair machine learning algorithms increasingly important. In this work we address the question: Is it possible to create fair summaries for massive datasets?
To this end, we develop the first streaming approximation algorithms for submodular maximization under fairness constraints, for both monotone and non-monotone functions. We validate our findings empirically on exemplar-based clustering, movie recommendation, DPP-based summarization, and maximum coverage in social networks, showing that fairness constraints do not significantly impact utility.
**********
Smoothed Geometry for Robust Attribution
Zifan Wang, Haofan Wang, Shakul Ramkumar, Piotr Mardziel, Matt Fredrikson, Anupam Datta
Feature attributions are a popular tool for explaining the behavior of Deep Neural Networks (DNNs), but have recently been shown to be vulnerable to attacks that produce divergent explanations for nearby inputs.
This lack of robustness is especially problematic in high-stakes applications where adversarially-manipulated explanations could impair safety and trustworthiness.
Building on a geometric understanding of these attacks presented in recent work, we identify Lipschitz continuity conditions on models' gradient that lead to robust gradient-based attributions, and observe that smoothness may also be related to the ability of an attack to transfer across multiple attribution methods.
To mitigate these attacks in practice, we propose an inexpensive regularization method that promotes these conditions in DNNs, as well as a stochastic smoothing technique that does not require re-training.
Our experiments on a range of image models demonstrate that both of these mitigations consistently improve attribution robustness, and confirm the role that smooth geometry plays in these attacks on real, large-scale models.
**********
Fast Adversarial Robustness Certification of Nearest Prototype Classifiers for Arbitrary Seminorms
Sascha Saralajew, Lars Holdijk, Thomas Villmann
Methods for adversarial robustness certification aim to provide an upper bound on the test error of a classifier under adversarial manipulation of its input. Current certification methods are computationally expensive and limited to attacks that optimize the manipulation with respect to a norm. We overcome these limitations by investigating the robustness properties of Nearest Prototype Classifiers (NPCs) like learning vector quantization and large margin nearest neighbor. For this purpose, we study the hypothesis margin. We prove that if NPCs use a dissimilarity measure induced by a seminorm, the hypothesis margin is a tight lower bound on the size of adversarial attacks and can be calculated in constant time—this provides the first adversarial robustness certificate calculable in reasonable time. Finally, we show that each NPC trained by a triplet loss maximizes the hypothesis margin and is therefore optimized for adversarial robustness. In the presented evaluation, we demonstrate that NPCs optimized for adversarial robustness are competitive with state-of-the-art methods and set a new benchmark with respect to computational complexity for robustness certification.
**********
Multi-agent active perception with prediction rewards
Mikko Lauri, Frans Oliehoek
Multi-agent active perception is a task where a team of agents cooperatively gathers observations to compute a joint estimate of a hidden variable. The task is decentralized and the joint estimate can only be computed after the task ends by

fusing observations of all agents. The objective is to maximize the accuracy of the estimate. The accuracy is quantified by a centralized prediction reward determined by a centralized decision-maker who perceives the observations gathered by all agents after the task ends. In this paper, we model multi-agent active perception as a decentralized partially observable Markov decision process (Dec-POMDP) with a convex centralized prediction reward. We prove that by introducing individual prediction actions for each agent, the problem is converted into a standard Dec-POMDP with a decentralized prediction reward. The loss due to decentralization is bounded, and we give a sufficient condition for when it is zero. Our results allow application of any Dec-POMDP solution algorithm to multi-agent active perception problems, and enable planning to reduce uncertainty without explicit computation of joint estimates. We demonstrate the empirical usefulness of our results by applying a standard Dec-POMDP algorithm to multi-agent active perception problems, showing increased scalability in the planning horizon.

**********

A Local Temporal Difference Code for Distributional Reinforcement Learning

Pablo Tano, Peter Dayan, Alexandre Pouget

Recent theoretical and experimental results suggest that the dopamine system implements distributional temporal difference backups, allowing learning of the entire distributions of the long-run values of states rather than just their expected values. However, the distributional codes explored so far rely on a complex imputation step which crucially relies on spatial non-locality: in order to compute reward prediction errors, units must know not only their own state but also the states of the other units. It is far from clear how these steps could be implemented in realistic neural circuits. Here, we introduce the Laplace code: a local temporal difference code for distributional reinforcement learning that is representationally powerful and computationally straightforward. The code decomposes value distributions and prediction errors across three separated dimensions: reward magnitude (related to distributional quantiles), temporal discounting (related to the Laplace transform of future rewards) and time horizon (related to eligibility traces). Besides lending itself to a local learning rule, the decomposition recovers the temporal evolution of the immediate reward distribution, indicating all possible rewards at all future times. This increases representational capacity and allows for temporally-flexible computations that immediately adjust to changing horizons or discount factors.

**********

Learning with Optimized Random Features: Exponential Speedup by Quantum Machine Learning without Sparsity and Low-Rank Assumptions

Hayata Yamasaki, Sathyawageeswar Subramanian, Sho Sonoda, Masato Koashi

Kernel methods augmented with random features give scalable algorithms for learning from big data. But it has been computationally hard to sample random features according to a probability distribution that is optimized for the data, so as to minimize the required number of features for achieving the learning to a desired accuracy. Here, we develop a quantum algorithm for sampling from this optimized distribution over features, in runtime O(D) that is linear in the dimension D of the input data. Our algorithm achieves an exponential speedup in D compared to any known classical algorithm for this sampling task. In contrast to existing quantum machine learning algorithms, our algorithm circumvents sparsity and low-rank assumptions and thus has wide applicability. We also show that the sampled features can be combined with regression by stochastic gradient descent to achieve the learning without canceling out our exponential speedup. Our algorithm based on sampling optimized random features leads to an accelerated framework for machine learning that takes advantage of quantum computers.

**********

CaSPR: Learning Canonical Spatiotemporal Point Cloud Representations

Davis Rempe, Tolga Birdal, Yongheng Zhao, Zan Gojcic, Srinath Sridhar, Leonidas J. Guibas

We propose CaSPR, a method to learn object-centric Canonical Spatiotemporal Point Cloud Representations of dynamically moving or evolving objects. Our goal is to enable information aggregation over time and the interrogation of object state

at any spatiotemporal neighborhood in the past, observed or not. Different from previous work, CaSPR learns representations that support spacetime continuity, are robust to variable and irregularly spacetime-sampled point clouds, and generalize to unseen object instances. Our approach divides the problem into two subtasks. First, we explicitly encode time by mapping an input point cloud sequence to a spatiotemporally-canonicalized object space. We then leverage this canonicalization to learn a spatiotemporal latent representation using neural ordinary differential equations and a generative model of dynamically evolving shapes using continuous normalizing flows. We demonstrate the effectiveness of our method on several applications including shape reconstruction, camera pose estimation, continuous spatiotemporal sequence reconstruction, and correspondence estimation from irregularly or intermittently sampled observations.

**********

Deep Automodulators

Ari Heljakka, Yuxin Hou, Juho Kannala, Arno Solin

We introduce a new category of generative autoencoders called automodulators. These networks can faithfully reproduce individual real-world input images like regular autoencoders, but also generate a fused sample from an arbitrary combination of several such images, allowing instantaneous "style-mixing" and other new applications. An automodulator decouples the data flow of decoder operations from statistical properties thereof and uses the latent vector to modulate the former by the latter, with a principled approach for mutual disentanglement of decoder layers. Prior work has explored similar decoder architecture with GANs, but their focus has been on random sampling. A corresponding autoencoder could operate on real input images. For the first time, we show how to train such a general-purpose model with sharp outputs in high resolution, using novel training techniques, demonstrated on four image data sets. Besides style-mixing, we show state-of-the-art results in autoencoder comparison, and visual image quality nearly indistinguishable from state-of-the-art GANs. We expect the automodulator variants to become a useful building block for image applications and other data domains.

**********

Convolutional Tensor-Train LSTM for Spatio-Temporal Learning

Jiahao Su, Wonmin Byeon, Jean Kossaifi, Furong Huang, Jan Kautz, Anima Anandkumar

Learning from spatio-temporal data has numerous applications such as human-behavior analysis, object tracking, video compression, and physics simulation. However, existing methods still perform poorly on challenging video tasks such as long-term forecasting. This is because these kinds of challenging tasks require learning long-term spatio-temporal correlations in the video sequence. In this paper, we propose a higher-order convolutional LSTM model that can efficiently learn these correlations, along with a succinct representations of the history. This is accomplished through a novel tensor train module that performs prediction by combining convolutional features across time. To make this feasible in terms of computation and memory requirements, we propose a novel convolutional tensor-train decomposition of the higher-order model. This decomposition reduces the model complexity by jointly approximating a sequence of convolutional kernels as a low-rank tensor-train factorization. As a result, our model outperforms existing approaches, but uses only a fraction of parameters, including the baseline models. Our results achieve state-of-the-art performance in a wide range of applications and datasets, including the multi-steps video prediction on the Moving-MNIST-2 and KTH action datasets as well as early activity recognition on the Something-Something V2 dataset.

**********

The Potts-Ising model for discrete multivariate data

Zahra Razaee, Arash Amini

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

**********

Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech
Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S. Turek, Alexander Huth
Natural language contains information at multiple timescales. To understand how the human brain represents this information, one approach is to build encoding models that predict fMRI responses to natural language using representations extracted from neural network language models (LMs). However, these LM-derived representations do not explicitly separate information at different timescales, making it difficult to interpret the encoding models. In this work we construct interpretable multi-timescale representations by forcing individual units in an LSTM LM to integrate information over specific temporal scales. This allows us to explicitly and directly map the timescale of information encoded by each individual fMRI voxel. Further, the standard fMRI encoding procedure does not account for varying temporal properties in the encoding features. We modify the procedure so that it can capture both short- and long-timescale information. This approach outperforms other encoding models, particularly for voxels that represent long-timescale information. It also provides a finer-grained map of timescale information in the human language pathway. This serves as a framework for future work investigating temporal hierarchies across artificial and biological language systems.
**********
Group-Fair Online Allocation in Continuous Time
Semih Cayci, Swati Gupta, Atilla Eryilmaz
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Decentralized TD Tracking with Linear Function Approximation and its Finite-Time Analysis
Gang Wang, Songtao Lu, Georgios Giannakis, Gerald Tesauro, Jian Sun
The present contribution deals with decentralized policy evaluation in multi-agent Markov decision processes using temporal-difference (TD) methods with linear function approximation for scalability. The agents cooperate to estimate the value function of such a process by observing continual state transitions of a shared environment over the graph of interconnected nodes (agents), along with locally private rewards. Different from existing consensus-type TD algorithms, the approach here develops a simple decentralized TD tracker by wedding TD learning with gradient tracking techniques. The non-asymptotic properties of the novel TD tracker are established for both independent and identically distributed (i.i.d.) as well as Markovian transitions through a unifying multistep Lyapunov analysis. In contrast to the prior art, the novel algorithm forgoes the limiting error bounds on the number of agents, which endows it with performance comparable to that of centralized TD methods that are the sharpest known to date.
**********
Understanding Gradient Clipping in Private SGD: A Geometric Perspective
Xiangyi Chen, Steven Z. Wu, Mingyi Hong
Deep learning models are increasingly popular in many machine learning applications where the training data may contain sensitive information. To provide formal and rigorous privacy guarantee, many learning systems now incorporate differential privacy by training their models with (differentially) private SGD. A key step in each private SGD update is gradient clipping that shrinks the gradient of an individual example whenever its l2 norm exceeds a certain threshold. We first demonstrate how gradient clipping can prevent SGD from converging to a stationary point. We then provide a theoretical analysis on private SGD with gradient clipping. Our analysis fully characterizes the clipping bias on the gradient norm, which can be upper bounded by the Wasserstein distance between the gradient distribution and a geometrically symmetric distribution. Our empirical evaluation further suggests that the gradient distributions along the trajectory of private

SGD indeed exhibit such symmetric structure. Together, our results provide an explanation why private SGD with gradient clipping remains effective in practice despite its potential clipping bias. Finally, we develop a new perturbation-based technique that can provably correct the clipping bias even for instances with highly asymmetric gradient distributions.
**********
O(n) Connections are Expressive Enough: Universal Approximability of Sparse Transformers
Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, 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-authors prior to requesting a name change in the electronic proceedings.
**********
Identifying signal and noise structure in neural population activity with Gaussian process factor models
Stephen Keeley, Mikio Aoi, Yiyi Yu, Spencer Smith, Jonathan W. Pillow
Neural datasets often contain measurements of neural activity across multiple trials of a repeated stimulus or behavior. An important problem in the analysis of such datasets is to characterize systematic aspects of neural activity that carry information about the repeated stimulus or behavior of interest, which can be considered signal'', and to separate them from the trial-to-trial fluctuations in activity that are not time-locked to the stimulus, which for purposes of such analyses can be considerednoise''. Gaussian Process factor models provide a powerful tool for identifying shared structure in high-dimensional neural data. However, they have not yet been adapted to the problem of characterizing signal and noise in multi-trial datasets. Here we address this shortcoming by proposing ``signal-noise'' Poisson-spiking Gaussian Process Factor Analysis (SNP-GPFA), a flexible latent variable model that resolves signal and noise latent structure in neural population spiking activity. To learn the parameters of our model, we introduce a Fourier-domain black box variational inference method that quickly identifies smooth latent structure. The resulting model reliably uncovers latent signal and trial-to-trial noise-related fluctuations in large-scale recordings. We use this model to show that in monkey V1, noise fluctuations perturb neural activity within a subspace orthogonal to signal activity, suggesting that trial-by-trial noise does not interfere with signal representations. Finally, we extend the model to capture statistical dependencies across brain regions in multi-region data. We show that in mouse visual cortex, models with shared noise across brain regions out-perform models with independent per-region noise.
**********
Equivariant Networks for Hierarchical Structures
Renhao Wang, Marjan Albooyeh, Siamak Ravanbakhsh
While using invariant and equivariant maps, it is possible to apply deep learning to a range of primitive data structures, a formalism for dealing with hierarchy is lacking. This is a significant issue because many practical structures are hierarchies of simple building blocks; some examples include sequences of sets, graphs of graphs, or multiresolution images. Observing that the symmetry of a hierarchical structure is the ``wreath product'' of symmetries of the building blocks, we express the equivariant map for the hierarchy using an intuitive combination of the equivariant linear layers of the building blocks. More generally, we show that any equivariant map for the hierarchy has this form. To demonstrate the effectiveness of this approach to model design, we consider its application in the semantic segmentation of point-cloud data. By voxelizing the point cloud, we impose a hierarchy of translation and permutation symmetries on the data and report state-of-the-art on {semantic3d}, {s3dis}, and {vkitti}, that include some of the largest real-world point-cloud benchmarks.
**********
MinMax Methods for Optimal Transport and Beyond: Regularization, Approximation and Numerics

Luca De Gennaro Aquino, Stephan Eckstein
We study MinMax solution methods for a general class of optimization problems related to (and including) optimal transport. Theoretically, the focus is on fitting a large class of problems into a single MinMax framework and generalizing regularization techniques known from classical optimal transport. We show that regularization techniques justify the utilization of neural networks to solve such problems by proving approximation theorems and illustrating fundamental issues if no regularization is used. We further study the relation to the literature on generative adversarial nets, and analyze which algorithmic techniques used therein are particularly suitable to the class of problems studied in this paper. Several numerical experiments showcase the generality of the setting and highlight which theoretical insights are most beneficial in practice.
**********
A Discrete Variational Recurrent Topic Model without the Reparametrization Trick
Mehdi Rezaee, Francis Ferraro
We show how to learn a neural topic model with discrete random variables---one that explicitly models each word's assigned topic---using neural variational inference that does not rely on stochastic backpropagation to handle the discrete variables. The model we utilize combines the expressive power of neural methods for representing sequences of text with the topic model's ability to capture global, thematic coherence. Using neural variational inference, we show improved perplexity and document understanding across multiple corpora. We examine the effect of prior parameters both on the model and variational parameters, and demonstrate how our approach can compete and surpass a popular topic model implementation on an automatic measure of topic quality.
**********
Transferable Graph Optimizers for ML Compilers
Yanqi Zhou, Sudip Roy, Amirali Abdolrashidi, Daniel Wong, Peter Ma, Qiumin Xu, Hanxiao Liu, Phitchaya Phothilimtha, Shen Wang, Anna Goldie, Azalia Mirhoseini, James Laudon
Most compilers for machine learning (ML) frameworks need to solve many correlated optimization problems to generate efficient machine code. Current ML compilers rely on heuristics based algorithms to solve these optimization problems one at a time. However, this approach is not only hard to maintain but often leads to sub-optimal solutions especially for newer model architectures. Existing learning based approaches in the literature are sample inefficient, tackle a single optimization problem, and do not generalize to unseen graphs making them infeasible to be deployed in practice. To address these limitations, we propose an end-to-end, transferable deep reinforcement learning method for computational graph optimization (GO), based on a scalable sequential attention mechanism over an inductive graph neural network. GO generates decisions on the entire graph rather than on each individual node autoregressively, drastically speeding up the search compared to prior methods. Moreover, we propose recurrent attention layers to jointly optimize dependent graph optimization tasks and demonstrate 33%-60% speedup on three graph optimization tasks compared to TensorFlow default optimization. On a diverse set of representative graphs consisting of up to 80,000 nodes, including Inception-v3, Transformer-XL, and WaveNet, GO achieves on average 21% improvement over human experts and 18% improvement over the prior state of the art with 15x faster convergence, on a device placement task evaluated in real systems.
**********
Learning with Operator-valued Kernels in Reproducing Kernel Krein Spaces
Akash Saha, Balamurugan Palaniappan
Operator-valued kernels have shown promise in supervised learning problems with functional inputs and functional outputs. The crucial (and possibly restrictive) assumption of positive definiteness of operator-valued kernels has been instrumental in developing efficient algorithms. In this work, we consider operator-valued kernels which might not be necessarily positive definite. To tackle the indefiniteness of operator-valued kernels, we harness the machinery of Reproducing Kernel Krein Spaces (RKKS) of function-valued functions. A representer theorem is

illustrated which yields a suitable loss stabilization problem for supervised learning with function-valued inputs and outputs. Analysis of generalization properties of the proposed framework is given. An iterative Operator based Minimum Residual (OpMINRES) algorithm is proposed for solving the loss stabilization problem. Experiments with indefinite operator-valued kernels on synthetic and real data sets demonstrate the utility of the proposed approach.
**********

Learning Bounds for Risk-sensitive Learning
Jaeho Lee, Sejun Park, Jinwoo Shin
In risk-sensitive learning, one aims to find a hypothesis that minimizes a risk-averse (or risk-seeking) measure of loss, instead of the standard expected loss. In this paper, we propose to study the generalization properties of risk-sensitive learning schemes whose optimand is described via optimized certainty equivalents (OCE): our general scheme can handle various known risks, e.g., the entropic risk, mean-variance, and conditional value-at-risk, as special cases. We provide two learning bounds on the performance of empirical OCE minimizer. The first result gives an OCE guarantee based on the Rademacher average of the hypothesis space, which generalizes and improves existing results on the expected loss and the conditional value-at-risk. The second result, based on a novel variance-based characterization of OCE, gives an expected loss guarantee with a suppressed dependence on the smoothness of the selected OCE. Finally, we demonstrate the practical implications of the proposed bounds via exploratory experiments on neural networks.
**********

Simplifying Hamiltonian and Lagrangian Neural Networks via Explicit Constraints
Marc Finzi, Ke Alexander Wang, Andrew G. Wilson
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency
Robert Geirhos, Kristof Meding, Felix A. Wichmann
A central problem in cognitive science and behavioural neuroscience as well as in machine learning and artificial intelligence research is to ascertain whether two or more decision makers---be they brains or algorithms---use the same strategy. Accuracy alone cannot distinguish between strategies: two systems may achieve similar accuracy with very different strategies. The need to differentiate beyond accuracy is particularly pressing if two systems are at or near ceiling performance, like Convolutional Neural Networks (CNNs) and humans on visual object recognition.
Here we introduce trial-by-trial error consistency, a quantitative analysis for measuring whether two decision making systems systematically make errors on the same inputs. Making consistent errors on a trial-by-trial basis is a necessary condition if we want to ascertain similar processing strategies between decision makers. Our analysis is applicable to compare algorithms with algorithms, humans with humans, and algorithms with humans.
When applying error consistency to visual object recognition we obtain three main findings: (1.) Irrespective of architecture, CNNs are remarkably consistent with one another. (2.) The consistency between CNNs and human observers, however, is little above what can be expected by chance alone---indicating that humans and CNNs are likely implementing very different strategies. (3.) CORnet-S, a recurrent model termed the "current best model of the primate ventral visual stream", fails to capture essential characteristics of human behavioural data and behaves essentially like a standard purely feedforward ResNet-50 in our analysis; highlighting that certain behavioural failure cases are not limited to feedforward models. Taken together, error consistency analysis suggests that the strategies used by human and machine vision are still very different---but we envision our general-purpose error consistency analysis to serve as a fruitful tool for quanti

fying future progress.
**********

Provably Efficient Reinforcement Learning with Kernel and Neural Function Approx imations
Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, Michael Jordan
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-auth ors prior to requesting a name change in the electronic proceedings.
**********

Constant-Expansion Suffices for Compressed Sensing with Generative Priors
Constantinos Daskalakis, Dhruv Rohatgi, Emmanouil Zampetakis
Generative neural networks have been empirically found very promising in providi ng effective structural priors for compressed sensing, since they can be trained to span low-dimensional data manifolds in high-dimensional signal spaces. Despi te the non-convexity of the resulting optimization problem, it has also been sho wn theoretically that, for neural networks with random Gaussian weights, a signa l in the range of the network can be efficiently, approximately recovered from a few noisy measurements. However, a major bottleneck of these theoretical guaran tees is a network \emph{expansivity} condition: that each layer of the neural ne twork must be larger than the previous by a logarithmic factor. Our main contrib ution is to break this strong expansivity assumption, showing that \emph{constan t} expansivity suffices to get efficient recovery algorithms, besides it also be ing information-theoretically necessary. To overcome the theoretical bottleneck in existing approaches we prove a novel uniform concentration theorem for random functions that might not be Lipschitz but satisfy a relaxed notion which we cal l ``pseudo-Lipschitzness.'' Using this theorem we can show that a matrix concent ration inequality known as the \emph{Weight Distribution Condition (WDC)}, which was previously only known to hold for Gaussian matrices with logarithmic aspect ratio, in fact holds for constant aspect ratios too. Since WDC is a fundamental matrix concentration inequality in the heart of all existing theoretical guaran tees on this problem, our tighter bound immediately yields improvements in all k nown results in the literature on compressed sensing with deep generative priors , including one-bit recovery, phase retrieval, and more.
**********

RANet: Region Attention Network for Semantic Segmentation
Dingguo Shen, Yuanfeng Ji, Ping Li, Yi Wang, Di Lin
Recent semantic segmentation methods model the relationship between pixels to co nstruct the contextual representations. In this paper, we introduce the \emph{Re gion Attention Network} (RANet), a novel attention network for modeling the rela tionship between object regions. RANet divides the image into object regions, wh ere we select representative information. In contrast to the previous methods, R ANet configures the information pathways between the pixels in different regions , enabling the region interaction to exchange the regional context for enhancing all of the pixels in the image. We train the construction of object regions, th e selection of the representative regional contents, the configuration of inform ation pathways and the context exchange between pixels, jointly, to improve the segmentation accuracy. We extensively evaluate our method on the challenging seg mentation benchmarks, demonstrating that RANet effectively helps to achieve the state-of-the-art results.
**********

A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent
Zhenyu Liao, Romain Couillet, Michael W. Mahoney
Requests for name changes in the electronic proceedings will be accepted with 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 sparse codes from compressed representations with biologically plausibl

e local wiring constraints
Kion Fallah, Adam Willats, Ninghao Liu, Christopher Rozell
Sparse coding is an important method for unsupervised learning of task-independe nt features in theoretical neuroscience models of neural coding. While a number of algorithms exist to learn these representations from the statistics of a data set, they largely ignore the information bottlenecks present in fiber pathways c onnecting cortical areas. For example, the visual pathway has many fewer neurons transmitting visual information to cortex than the number of photoreceptors. B oth empirical and analytic results have recently shown that sparse representatio ns can be learned effectively after performing dimensionality reduction with ran domized linear operators, producing latent coefficients that preserve informatio n. Unfortunately,current proposals for sparse coding in the compressed space re quire a centralized compression process (i.e., dense random matrix) that is biol ogically unrealistic due to local wiring constraints observed in neural circuits . The main contribution of this paper is to leverage recent results on structure d random matrices to propose a theoretical neuroscience model of randomized proj ections for communication between cortical areas that is consistent with the loc al wiring constraints observed in neuroanatomy. We show analytically and empiric ally that unsupervised learning of sparse representations can be performed in th e compressed space despite significant local wiring constraints in compression m atrices of varying forms (corresponding to different local wiring patterns). Our analysis verifies that even with significant local wiring constraints, the lear ned representations remain qualitatively similar,have similar quantitative perfo rmance in both training and generalization error, and are consistent across many measures with measured macaque V1 receptive fields.
**********
Self-Imitation Learning via Generalized Lower Bound Q-learning
Yunhao Tang
Self-imitation learning motivated by lower-bound Q-learning is a novel and effec tive approach for off-policy learning. In this work, we propose a n-step lower b ound which generalizes the original return-based lower-bound Q-learning, and int roduce a new family of self-imitation learning algorithms. To provide a formal m otivation for the potential performance gains provided by self-imitation learnin g, we show that n-step lower bound Q-learning achieves a trade-off between fixed point bias and contraction rate, drawing close connections to the popular uncor rected n-step Q-learning. We finally show that n-step lower bound Q-learning is a more robust alternative to return-based self-imitation learning and uncorrecte d n-step, over a wide range of benchmark tasks.
**********
Private Learning of Halfspaces: Simplifying the Construction and Reducing the Sa mple Complexity
Haim Kaplan, Yishay Mansour, Uri Stemmer, Eliad Tsfadia
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Directional Pruning of Deep Neural Networks
Shih-Kang Chao, Zhanyu Wang, Yue Xing, Guang Cheng
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Smoothly Bounding User Contributions in Differential Privacy
Alessandro Epasto, Mohammad Mahdian, Jieming Mao, Vahab Mirrokni, Lijie Ren
A differentially private algorithm guarantees that the input of a single user wo n’t significantly change the output distribution of the algorithm. When a user c ontributes more data points, more information can be collected to improve the al gorithm’s performance. But at the same time, more noise might need to be added t

o the algorithm in order to keep the algorithm differentially private and this m
ight hurt the algorithm’s performance. Amin et al. (2019) initiates the study on
 bounding user contributions and proposes a very natural algorithm which limits
the number of samples each user can contribute by a threshold.
**********

Accelerating Training of Transformer-Based Language Models with Progressive Laye
r Dropping
Minjia Zhang, Yuxiong He
Recently, Transformer-based language models have demonstrated remarkable perform
ance across many NLP domains. However, the unsupervised pre-training step of the
se models suffers from unbearable overall computational expenses. Current method
s for accelerating the pre-training either rely on massive parallelism with adva
nced hardware or are not applicable to language models.
**********

Online Planning with Lookahead Policies
Yonathan Efroni, Mohammad Ghavamzadeh, Shie Mannor
Requests for name changes in the electronic proceedings will be accepted with 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 Deep Attribution Priors Based On Prior Knowledge
Ethan Weinberger, Joseph Janizek, Su-In Lee
Feature attribution methods, which explain an individual prediction made by a mo
del as a sum of attributions for each input feature, are an essential tool for u
nderstanding the behavior of complex deep learning models.  However, ensuring th
at models produce meaningful explanations, rather than ones that rely on noise,
is not straightforward. Exacerbating this problem is the fact that attribution m
ethods do not provide insight as to why features are assigned their attribution
values, leading to explanations that are difficult to interpret.  In real-world
problems we often have sets of additional information for each feature that are
predictive of that feature's importance to the task at hand. Here, we propose th
e deep attribution prior (DAPr) framework to exploit such information to overcom
e the limitations of attribution methods.  Our framework jointly learns a relati
onship between prior information and feature importance, as well as biases model
s to have explanations that rely on features predicted to be important.  We find
 that our framework both results in networks that generalize better to out of sa
mple data and admits new methods for interpreting model behavior.
**********

Using noise to probe recurrent neural network structure and prune synapses
Eli Moore, Rishidev Chaudhuri
Many networks in the brain are sparsely connected, and the brain eliminates syna
pses during development and learning. How could the brain decide which synapses
to prune? In a recurrent network, determining the importance of a synapse betwee
n two neurons is a difficult computational problem, depending on the role that b
oth neurons play and on all possible pathways of information flow between them.
Noise is ubiquitous in neural systems, and often considered an irritant to be ov
ercome. Here we suggest that noise could play a functional role in synaptic prun
ing, allowing the brain to probe network structure and determine which synapses
are redundant. We construct a simple, local, unsupervised plasticity rule that e
ither strengthens or prunes synapses using only synaptic weight and the noise-dr
iven covariance of the neighboring neurons. For a subset of linear and rectified
-linear networks, we prove that this rule preserves the spectrum of the original
 matrix and hence preserves network dynamics even when the fraction of pruned sy
napses asymptotically approaches 1. The plasticity rule is biologically-plausibl
e and may suggest a new role for noise in neural computation.
**********

NanoFlow: Scalable Normalizing Flows with Sublinear Parameter Complexity
Sang-gil Lee, Sungwon Kim, Sungroh Yoon
Normalizing flows (NFs) have become a prominent method for deep generative model

s that allow for an analytic probability density estimation and efficient synthesis. However, a flow-based network is considered to be inefficient in parameter complexity because of reduced expressiveness of bijective mapping, which renders the models unfeasibly expensive in terms of parameters. We present an alternative parameterization scheme called NanoFlow, which uses a single neural density estimator to model multiple transformation stages. Hence, we propose an efficient parameter decomposition method and the concept of flow indication embedding, which are key missing components that enable density estimation from a single neural network. Experiments performed on audio and image models confirm that our method provides a new parameter-efficient solution for scalable NFs with significant sublinear parameter complexity.

**********

Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge

Chaoyang He, Murali Annavaram, Salman Avestimehr

Scaling up the convolutional neural network (CNN) size (e.g., width, depth, etc.) is known to effectively improve model accuracy. However, the large model size impedes training on resource-constrained edge devices. For instance, federated learning (FL) may place undue burden on the compute capability of edge nodes, even though there is a strong practical need for FL due to its privacy and confidentiality properties. To address the resource-constrained reality of edge devices, we reformulate FL as a group knowledge transfer training algorithm, called FedGKT. FedGKT designs a variant of the alternating minimization approach to train small CNNs on edge nodes and periodically transfer their knowledge by knowledge distillation to a large server-side CNN. FedGKT consolidates several advantages into a single framework: reduced demand for edge computation, lower communication bandwidth for large CNNs, and asynchronous training, all while maintaining model accuracy comparable to FedAvg. We train CNNs designed based on ResNet-56 and ResNet-110 using three distinct datasets (CIFAR-10, CIFAR-100, and CINIC-10) and their non-IID variants. Our results show that FedGKT can obtain comparable or even slightly higher accuracy than FedAvg. More importantly, FedGKT makes edge training affordable. Compared to the edge training using FedAvg, FedGKT demands 9 to 17 times less computational power (FLOPs) on edge devices and requires 54 to 105 times fewer parameters in the edge CNN. Our source code is released at FedML (https://fedml.ai).

**********

Neural FFTs for Universal Texture Image Synthesis

Morteza Mardani, Guilin Liu, Aysegul Dundar, Shiqiu Liu, Andrew Tao, Bryan Catanzaro

Synthesizing larger texture images from a smaller exemplar is an important task in graphics and vision. The conventional CNNs, recently adopted for synthesis, require to train and test on the same set of images and fail to generalize to unseen images. This is mainly because those CNNs fully rely on convolutional and upsampling layers that operate locally and not suitable for a task as global as texture synthesis. In this work, inspired by the repetitive nature of texture patterns, we find that texture synthesis can be viewed as (local) \textit{upsampling} in the Fast Fourier Transform (FFT) domain. However, FFT of natural images exhibits high dynamic range and lacks local correlations. Therefore, to train CNNs we design a framework to perform FFT upsampling in feature space using deformable convolutions. Such design allows our framework to generalize to unseen images, and synthesize textures in a single pass. Extensive evaluations confirm that our method achieves state-of-the-art performance both quantitatively and qualitatively.

**********

Graph Cross Networks with Vertex Infomax Pooling

Maosen Li, Siheng Chen, Ya Zhang, Ivor Tsang

We propose a novel graph cross network (GXN) to achieve comprehensive feature learning from multiple scales of a graph. Based on trainable hierarchical representations of a graph, GXN enables the interchange of intermediate features across scales to promote information flow. Two key ingredients of GXN include a novel vertex infomax pooling (VIPool), which creates multiscale graphs in a trainable m

anner, and a novel feature-crossing layer, enabling feature interchange across s cales.  The proposed VIPool selects the most informative subset of vertices base d on the neural estimation of mutual information between vertex features and nei ghborhood features. The intuition behind is that a vertex is informative when it  can maximally reflect its neighboring information. The proposed feature-crossin g layer fuses intermediate features between two scales for mutual enhancement by  improving information flow and enriching multiscale features at hidden layers. The cross shape of feature-crossing layer distinguishes GXN from many other mult iscale architectures. Experimental results show that the proposed GXN improves t he classification accuracy  by 2.12% and 1.15% on average for graph classificati on and vertex classification, respectively. Based on the same network, the propo sed VIPool consistently outperforms other graph-pooling methods.
**********

Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms

Hilal Asi, John C. Duchi

We study and provide instance-optimal algorithms in differential privacy by exte nding and approximating the inverse sensitivity mechanism. We provide two approx imation frameworks, one which only requires knowledge of local sensitivities, an d a gradient-based approximation for optimization problems, which are efficientl y computable for a broad class of functions. We complement our analysis with ins tance-specific lower bounds for vector-valued functions, which demonstrate that our mechanisms are (nearly) instance-optimal under certain assumptions and that minimax lower bounds may not provide an accurate estimate of the hardness of a p roblem in general: our algorithms can significantly outperform minimax bounds fo r well-behaved instances. Finally, we use our approximation framework to develop  private mechanisms for unbounded-range mean estimation, principal component ana lysis, and linear regression. For PCA, our mechanisms give an efficient (pure) d ifferentially private algorithm with near-optimal rates.
**********

Calibration of Shared Equilibria in General Sum Partially Observable Markov Game s

Nelson Vadori, Sumitra Ganesh, Prashant Reddy, Manuela Veloso

Training multi-agent systems (MAS) to achieve realistic equilibria gives us a us eful tool to understand and model real-world systems. We consider a general sum partially observable Markov game where agents of different types share a single policy network, conditioned on agent-specific information. This paper aims at i)  formally understanding equilibria reached by such agents, and ii) matching emer gent phenomena of such equilibria to real-world targets. Parameter sharing with decentralized execution has been introduced as an efficient way to train multipl e agents using a single policy network. However, the nature of resulting equilib ria reached by such agents has not been yet studied: we introduce the novel conc ept of Shared equilibrium as a symmetric pure Nash equilibrium of a certain Func tional Form Game (FFG) and prove convergence to the latter for a certain class o f games using self-play. In addition, it is important that such equilibria satis fy certain constraints so that MAS are calibrated to real world data for practic al use: we solve this problem by introducing a novel dual-Reinforcement Learning  based approach that fits emergent behaviors of agents in a Shared equilibrium t o externally-specified targets, and apply our methods to a n-player market examp le. We do so by calibrating parameters governing distributions of agent types ra ther than individual agents, which allows both behavior differentiation among ag ents and coherent scaling of the shared policy network to multiple agents.
**********

MOPO: Model-based Offline Policy Optimization

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine , Chelsea Finn, Tengyu Ma

Offline reinforcement learning (RL) refers to the problem of learning policies e ntirely from a batch of previously collected data. This problem setting is compe lling, because it offers the promise of utilizing large, diverse, previously col lected datasets to acquire policies without any costly or dangerous active explo

ration, but it is also exceptionally difficult, due to the distributional shift between the offline training data and the learned policy. While there has been significant progress in model-free offline RL, the most successful prior methods constrain the policy to the support of the data, precluding generalization to new states. In this paper, we observe that an existing model-based RL algorithm on its own already produces significant gains in the offline setting, as compared to model-free approaches, despite not being designed for this setting. However, although many standard model-based RL methods already estimate the uncertainty of their model, they do not by themselves provide a mechanism to avoid the issues associated with distributional shift in the offline setting. We therefore propose to modify existing model-based RL methods to address these issues by casting offline model-based RL into a penalized MDP framework. We theoretically show that, by using this penalized MDP, we are maximizing a lower bound of the return in the true MDP. Based on our theoretical results, we propose a new model-based offline RL algorithm that applies the variance of a Lipschitz-regularized model as a penalty to the reward function. We find that this algorithm outperforms both standard model-based RL methods and existing state-of-the-art model-free offline RL approaches on existing offline RL benchmarks, as well as two challenging continuous control tasks that require generalizing from data collected for a different task.

**********

Building powerful and equivariant graph neural networks with structural message-passing

Clément Vignac, Andreas Loukas, Pascal Frossard

Message-passing has proved to be an effective way to design graph neural networks, as it is able to leverage both permutation equivariance and an inductive bias towards learning local structures in order to achieve good generalization. However, current message-passing architectures have a limited representation power and fail to learn basic topological properties of graphs. We address this problem and propose a powerful and equivariant message-passing framework based on two ideas: first, we propagate a one-hot encoding of the nodes, in addition to the features, in order to learn a local context matrix around each node. This matrix contains rich local information about both features and topology and can eventually be pooled to build node representations. Second, we propose methods for the parametrization of the message and update functions that ensure permutation equivariance. Having a representation that is independent of the specific choice of the one-hot encoding permits inductive reasoning and leads to better generalization properties. Experimentally, our model can predict various graph topological properties on synthetic data more accurately than previous methods and achieves state-of-the-art results on molecular graph regression on the ZINC dataset.

**********

Efficient Model-Based Reinforcement Learning through Optimistic Policy Search and Planning

Sebastian Curi, Felix Berkenkamp, Andreas Krause

Model-based reinforcement learning algorithms with probabilistic dynamical models are amongst the most data-efficient learning methods. This is often attributed to their ability to distinguish between epistemic and aleatoric uncertainty. However, while most algorithms distinguish these two uncertainties for learning the model, they ignore it when optimizing the policy, which leads to greedy and insufficient exploration. At the same time, there are no practical solvers for optimistic exploration algorithms. In this paper, we propose a practical optimistic exploration algorithm (H-UCRL). H-UCRL reparameterizes the set of plausible models and hallucinates control directly on the epistemic uncertainty. By augmenting the input space with the hallucinated inputs, H-UCRL can be solved using standard greedy planners. Furthermore, we analyze H-UCRL and construct a general regret bound for well-calibrated models, which is provably sublinear in the case of Gaussian Process models. Based on this theoretical foundation, we show how optimistic exploration can be easily combined with state-of-the-art reinforcement learning algorithms and different probabilistic models. Our experiments demonstrate that optimistic exploration significantly speeds-up learning when there are pen

alties on actions, a setting that is notoriously difficult for existing model-ba
sed reinforcement learning algorithms.
**********

Practical Low-Rank Communication Compression in Decentralized Deep Learning
Thijs Vogels, Sai Praneeth Karimireddy, Martin Jaggi
Lossy gradient compression has become a practical tool to overcome the communica
tion bottleneck in centrally coordinated distributed training of machine learnin
g models. However, algorithms for decentralized training with compressed communi
cation over arbitrary connected networks have been more complicated, requiring a
dditional memory and hyperparameters. We introduce a simple algorithm that direc
tly compresses the model differences between neighboring workers using low-rank
linear compressors. We prove that our method does not require any additional hyp
erparameters, converges faster than prior methods, and is asymptotically indepen
dent of both the network and the compression. Inspired the PowerSGD algorithm fo
r centralized deep learning, we execute power iteration steps on model differenc
es to maximize the information transferred per bit. Out of the box, these compre
ssors perform on par with state-of-the-art tuned compression algorithms in a ser
ies of deep learning benchmarks.
**********

Mutual exclusivity as a challenge for deep neural networks
Kanishk Gandhi, Brenden M. Lake
Strong inductive biases allow children to learn in fast and adaptable ways. Chil
dren use the mutual exclusivity (ME) bias to help disambiguate how words map to
referents, assuming that if an object has one label then it does not need anothe
r. In this paper, we investigate whether or not vanilla neural architectures hav
e an ME bias, demonstrating that they lack this learning assumption. Moreover, w
e show that their inductive biases are poorly matched to lifelong learning formu
lations of classification and translation. We demonstrate that there is a compel
ling case for designing task-general neural networks that learn through mutual e
xclusivity, which remains an open challenge.
**********

3D Shape Reconstruction from Vision and Touch
Edward Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, J
itendra Malik, Michal Drozdzal
When a toddler is presented a new toy, their instinctual behaviour is to pick it
up and inspect it with their hand and eyes in tandem, clearly searching over it
s surface to properly understand what they are playing with. At any instance her
e, touch provides high fidelity localized information while vision provides comp
lementary global context. However, in 3D shape reconstruction, the complementary
fusion of visual and haptic modalities remains largely unexplored. In this pape
r, we study this problem and present an effective chart-based approach to multi-
modal shape understanding which encourages a similar fusion vision and touch inf
ormation. To do so, we introduce a dataset of simulated touch and vision signals
from the interaction between a robotic hand and a large array of 3D objects. Ou
r results show that (1) leveraging both vision and touch signals consistently im
proves single- modality baselines; (2) our approach outperforms alternative moda
lity fusion methods and strongly benefits from the proposed chart-based structur
e; (3) the reconstruction quality increases with the number of grasps provided;
and (4) the touch information not only enhances the reconstruction at the touch
site but also extrapolates to its local neighborhood.
**********

GradAug: A New Regularization Method for Deep Neural Networks
Taojiannan Yang, Sijie Zhu, Chen Chen
We propose a new regularization method to alleviate over-fitting in deep neural
networks. The key idea is utilizing randomly transformed training samples to reg
ularize a set of sub-networks, which are originated by sampling the width of the
original network, in the training process. As such, the proposed method introdu
ces self-guided disturbances to the raw gradients of the network and therefore i
s termed as Gradient Augmentation (GradAug). We demonstrate that GradAug can hel
p the network learn well-generalized and more diverse representations. Moreover,

it is easy to implement and can be applied to various structures and applicatio
ns. GradAug improves ResNet-50 to 78.79% on ImageNet classification, which is a
new state-of-the-art accuracy. By combining with CutMix, it further boosts the p
erformance to 79.67%, which outperforms an ensemble of advanced training tricks.
The generalization ability is evaluated on COCO object detection and instance s
egmentation where GradAug significantly surpasses other state-of-the-art methods
. GradAug is also robust to image distortions and FGSM adversarial attacks and i
s highly effective in low data regimes. Code is available at \url{https://github
.com/taoyang1122/GradAug}
**********

An Equivalence between Loss Functions and Non-Uniform Sampling in Experience Rep
lay
Scott Fujimoto, David Meger, Doina Precup
Prioritized Experience Replay (PER) is a deep reinforcement learning technique i
n which agents learn from transitions sampled with non-uniform probability propo
rtionate to their temporal-difference error. We show that any loss function eval
uated with non-uniformly sampled data can be transformed into another uniformly
sampled loss function with the same expected gradient. Surprisingly, we find in
some environments PER can be replaced entirely by this new loss function without
impact to empirical performance. Furthermore, this relationship suggests a new
branch of improvements to PER by correcting its uniformly sampled loss function
equivalent. We demonstrate the effectiveness of our proposed modifications to PE
R and the equivalent loss function in several MuJoCo and Atari environments.
**********

Learning Utilities and Equilibria in Non-Truthful Auctions
Hu Fu, Tao 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.
**********

Rational neural networks
Nicolas Boulle, Yuji Nakatsukasa, Alex Townsend
We consider neural networks with rational activation functions. The choice of th
e nonlinear activation function in deep learning architectures is crucial and he
avily impacts the performance of a neural network. We establish optimal bounds i
n terms of network complexity and prove that rational neural networks approximat
e smooth functions more efficiently than ReLU networks with exponentially smalle
r depth. The flexibility and smoothness of rational activation functions make th
em an attractive alternative to ReLU, as we demonstrate with numerical experimen
ts.
**********

DISK: Learning local features with policy gradient
Micha■ Tyszkiewicz, Pascal Fua, Eduard Trulls
Local feature frameworks are difficult to learn in an end-to-end fashion due to
the discreteness inherent to the selection and matching of sparse keypoints. We
introduce DISK (DIScrete Keypoints), a novel method that overcomes these obstacl
es by leveraging principles from Reinforcement Learning (RL), optimizing end-to-
end for a high number of correct feature matches. Our simple yet expressive prob
abilistic model lets us keep the training and inference regimes close, while mai
ntaining good enough convergence properties to reliably train from scratch. Our
features can be extracted very densely while remaining discriminative, challengi
ng commonly held assumptions about what constitutes a good keypoint, as showcase
d in Fig. 1, and deliver state-of-the-art results on three public benchmarks.
**********

Transfer Learning via $\ell_1$ Regularization
Masaaki Takada, Hironori Fujisawa
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
GOCor: Bringing Globally Optimized Correspondence Volumes into Your Neural Netwo
rk
Prune Truong, Martin Danelljan, Luc V. Gool, Radu Timofte
The feature correlation layer serves as a key neural network module in numerous
computer vision problems that involve dense correspondences between image pairs.
 It predicts a correspondence volume by evaluating dense scalar products between
 feature vectors extracted from pairs of locations in two images.
However, this point-to-point feature comparison is insufficient when disambiguat
ing multiple similar regions in an image, severely affecting the performance of
the end task.
We propose GOCor, a fully differentiable dense matching module, acting as a dire
ct replacement to the feature correlation layer.
The correspondence volume generated by our module is the result of an internal o
ptimization procedure that explicitly accounts for similar regions in the scene.
 Moreover, our approach is capable of effectively learning spatial matching prio
rs to resolve further matching ambiguities.
We analyze our GOCor module in extensive ablative experiments. When integrated i
nto state-of-the-art networks, our approach significantly outperforms the featur
e correlation layer for the tasks of geometric matching, optical flow, and dense
 semantic matching. The code and trained models will be made available at github
.com/PruneTruong/GOCor.
**********
Deep Inverse Q-learning with Constraints
Gabriel Kalweit, Maria Huegle, Moritz Werling, Joschka Boedecker
Popular Maximum Entropy Inverse Reinforcement Learning approaches require the co
mputation of expected state visitation frequencies for the optimal policy under
an estimate of the reward function. This usually requires intermediate value est
imation in the inner loop of the algorithm, slowing down convergence considerabl
y. In this work, we introduce a novel class of algorithms that only needs to sol
ve the MDP underlying the demonstrated behavior once to recover the expert polic
y. This is possible through a formulation that exploits a probabilistic behavior
 assumption for the demonstrations within the structure of Q-learning. We propos
e Inverse Action-value Iteration which is able to fully recover an underlying re
ward of an external agent in closed-form analytically. We further provide an acc
ompanying class of sampling-based variants which do not depend on a model of the
 environment. We show how to extend this class of algorithms to continuous state
-spaces via function approximation and how to estimate a corresponding action-va
lue function, leading to a policy as close as possible to the policy of the exte
rnal agent, while optionally satisfying a list of predefined hard constraints. W
e evaluate the resulting algorithms called Inverse Action-value Iteration, Inver
se Q-learning and Deep Inverse Q-learning on the Objectworld benchmark, showing
a speedup of up to several orders of magnitude compared to (Deep) Max-Entropy al
gorithms. We further apply Deep Constrained Inverse Q-learning on the task of le
arning autonomous lane-changes in the open-source simulator SUMO achieving compe
tent driving after training on data corresponding to 30 minutes of demonstration
s.
**********
Optimistic Dual Extrapolation for Coherent Non-monotone Variational Inequalities
Chaobing Song, Zhengyuan Zhou, Yichao Zhou, Yong Jiang, Yi 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.
**********
Prediction with Corrupted Expert Advice
Idan Amir, Idan Attias, Tomer Koren, Yishay Mansour, Roi Livni
We revisit the fundamental problem of prediction with expert advice, in a settin
g where the environment is benign and generates losses stochastically, but the f

eedback observed by the learner is subject to a moderate adversarial corruption. We prove that a variant of the classical Multiplicative Weights algorithm with decreasing step sizes achieves constant regret in this setting and performs optimally in a wide range of environments, regardless of the magnitude of the injected corruption. Our results reveal a surprising disparity between the often comparable Follow the Regularized Leader (FTRL) and Online Mirror Descent (OMD) frameworks: we show that for experts in the corrupted stochastic regime, the regret performance of OMD is in fact strictly inferior to that of FTRL.

**********

Human Parsing Based Texture Transfer from Single Image to 3D Human via Cross-View Consistency

Fang Zhao, Shengcai Liao, Kaihao Zhang, Ling Shao

This paper proposes a human parsing based texture transfer model via cross-view consistency learning to generate the texture of 3D human body from a single image. We use the semantic parsing of human body as input for providing both the shape and pose information to reduce the appearance variation of human image and preserve the spatial distribution of semantic parts. Meanwhile, in order to improve the prediction for textures of invisible parts, we explicitly enforce the consistency across different views of the same subject by exchanging the textures predicted by two views to render images during training. The perception loss and total variation regularization are optimized to maximize the similarity between rendered and input images, which does not necessitate extra 3D texture supervision. Experimental results on pedestrian images and fashion photos demonstrate that our method can produce higher quality textures with convincing details than other texture generation methods.

**********

Knowledge Augmented Deep Neural Networks for Joint Facial Expression and Action Unit Recognition

Zijun Cui, Tengfei Song, Yuru Wang, Qiang Ji

Facial expression and action units (AUs) represent two levels of descriptions of the facial behavior. Due to the underlying facial anatomy and the need to form a meaningful coherent expression, they are strongly correlated. This paper proposes to systematically capture their dependencies and incorporate them into a deep learning framework for joint facial expression recognition and action unit detection. Specifically, we first propose a constraint optimization method to encode the generic knowledge on expression-AUs probabilistic dependencies into a Bayesian Network (BN). The BN is then integrated into a deep learning framework as a weak supervision for an AU detection model. A data-driven facial expression recognition(FER) model is then constructed from data. Finally, the FER model and AU detection model are trained jointly to refine their learning. Evaluations on benchmark datasets demonstrate the effectiveness of the proposed knowledge integration in improving the performance of both the FER model and the AU detection model. The proposed AU detection model is demonstrated to be able to achieve competitive performance without AU annotations. Furthermore, the proposed Bayesian Network capturing the generic knowledge is demonstrated to generalize well to different datasets.

**********

Point process models for sequence detection in high-dimensional neural spike trains

Alex Williams, Anthony Degleris, Yixin Wang, Scott Linderman

Sparse sequences of neural spikes are posited to underlie aspects of working memory, motor production, and learning. Discovering these sequences in an unsupervised manner is a longstanding problem in statistical neuroscience. Promising recent work utilized a convolutive nonnegative matrix factorization model to tackle this challenge. However, this model requires spike times to be discretized, utilizes a sub-optimal least-squares criterion, and does not provide uncertainty estimates for model predictions or estimated parameters. We address each of these shortcomings by developing a point process model that characterizes fine-scale sequences at the level of individual spikes and represents sequence occurrences as a small number of marked events in continuous time. This ultra-sparse represent

ation of sequence events opens new possibilities for spike train modeling. For e xample, we introduce learnable time warping parameters to model sequences of var ying duration, which have been experimentally observed in neural circuits. We de monstrate these advantages on recordings from songbird higher vocal center and r odent hippocampus.
**********

Adversarial Attacks on Linear Contextual Bandits
Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Te ytaud, Alessandro Lazaric, Matteo Pirotta
Contextual bandit algorithms are applied in a wide range of domains, from advert ising to recommender systems, from clinical trials to education. In many of thes e domains, malicious agents may have incentives to force a bandit algorithm into a desired behavior For instance, an unscrupulous ad publisher may try to increa se their own revenue at the expense of the advertisers; a seller may want to inc rease the exposure of their products, or thwart a competitor’s advertising campa ign. In this paper, we study several attack scenarios and show that a malicious agent can force a linear contextual bandit algorithm to pull any desired arm T − o(T) times over a horizon of T steps, while applying adversarial modifications to either rewards or contexts with a cumulative cost that only grow logarithmica lly as O(log T). We also investigate the case when a malicious agent is interest ed in affecting the behavior of the bandit algorithm in a single context (e.g., a specific user). We first provide sufficient conditions for the feasibility of the attack and an efficient algorithm to perform an attack. We empirically valid ate the proposed approaches on synthetic and real-world datasets.
**********

Meta-Consolidation for Continual Learning
Joseph K J, Vineeth N Balasubramanian
The ability to continuously learn and adapt itself to new tasks, without losing grasp of already acquired knowledge is a hallmark of biological learning systems , which current deep learning systems fall short of. In this work, we present a novel methodology for continual learning called MERLIN: Meta-Consolidation for C ontinual Learning.
**********

Organizing recurrent network dynamics by task-computation to enable continual le arning
Lea Duncker, Laura Driscoll, Krishna V. Shenoy, Maneesh Sahani, David Sussillo
Biological systems face dynamic environments that require continual learning. It is not well understood how these systems balance the tension between flexibilit y for learning and robustness for memory of previous behaviors. Continual learni ng without catastrophic interference also remains a challenging problem in machi ne learning. Here, we develop a novel learning rule designed to minimize interfe rence between sequentially learned tasks in recurrent networks. Our learning rul e preserves network dynamics within activity-defined subspaces used for previous ly learned tasks. It encourages dynamics associated with new tasks that might ot herwise interfere to instead explore orthogonal subspaces, and it allows for reu se of previously established dynamical motifs where possible. Employing a set of tasks used in neuroscience, we demonstrate that our approach successfully elimi nates catastrophic interference and offers a substantial improvement over previo us continual learning algorithms. Using dynamical systems analysis, we show that networks trained using our approach can reuse similar dynamical structures acro ss similar tasks. This possibility for shared computation allows for faster lear ning during sequential training. Finally, we identify organizational differences that emerge when training tasks sequentially versus simultaneously.
**********

Lifelong Policy Gradient Learning of Factored Policies for Faster Training Witho ut Forgetting
Jorge Mendez, Boyu Wang, Eric Eaton
Policy gradient methods have shown success in learning control policies for high -dimensional dynamical systems. Their biggest downside is the amount of explorat ion they require before yielding high-performing policies. In a lifelong learnin

g setting, in which an agent is faced with multiple consecutive tasks over its l
ifetime, reusing information from previously seen tasks can substantially accele
rate the learning of new tasks. We provide a novel method for lifelong policy gr
adient learning that trains lifelong function approximators directly via policy
gradients, allowing the agent to benefit from accumulated knowledge throughout t
he entire training process. We show empirically that our algorithm learns faster
 and converges to better policies than single-task and lifelong learning baselin
es, and completely avoids catastrophic forgetting on a variety of challenging do
mains.
**********

Kernel Methods Through the Roof: Handling Billions of Points Efficiently
Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, Alessandro Rudi
Kernel methods provide an elegant and principled approach to nonparametric learn
ing, but so far could hardly be used in large scale problems, since naïve implem
entations scale poorly with data size.
Recent advances have shown the benefits of a number of algorithmic ideas, for ex
ample combining optimization, numerical linear algebra and random projections.
Here, we push these efforts further to develop and test a solver that takes full
 advantage of GPU hardware.
Towards this end, we designed a preconditioned gradient solver for kernel method
s  exploiting both GPU acceleration and parallelization with multiple GPUs, impl
ementing out-of-core variants of common linear algebra operations to guarantee o
ptimal hardware utilization.
Further, we optimize the numerical precision of different
operations and maximize efficiency of matrix-vector multiplications. As a result
we can experimentally show dramatic speedups on datasets with billions of points
,
while still guaranteeing state of the art performance.
Additionally, we make our software available as an easy to use library.
**********

Spike and slab variational Bayes for high dimensional logistic regression
Kolyan Ray, Botond Szabo, Gabriel Clara
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Maximum-Entropy Adversarial Data Augmentation for Improved Generalization and Ro
bustness
Long Zhao, Ting Liu, Xi Peng, Dimitris Metaxas
Adversarial data augmentation has shown promise for training robust deep neural
networks against unforeseen data shifts or corruptions. However, it is difficult
 to define heuristics to generate effective fictitious target distributions cont
aining "hard" adversarial perturbations that are largely different from the sour
ce distribution. In this paper, we propose a novel and effective regularization
term for adversarial data augmentation. We theoretically derive it from the info
rmation bottleneck principle, which results in a maximum-entropy formulation. In
tuitively, this regularization term encourages perturbing the underlying source
distribution to enlarge predictive uncertainty of the current model, so that the
 generated "hard" adversarial perturbations can improve the model robustness dur
ing training. Experimental results on three standard benchmarks demonstrate that
 our method consistently outperforms the existing state of the art by a statisti
cally significant margin.
**********

Fast geometric learning with symbolic matrices
Jean Feydy, Alexis Glaunès, Benjamin Charlier, Michael Bronstein
Geometric methods rely on tensors that can be encoded using a symbolic formula a
nd data arrays, such as kernel and distance matrices. We present an extension fo
r standard machine learning frameworks that provides comprehensive support for t
his abstraction on CPUs and GPUs: our toolbox combines a versatile, transparent

user interface with fast runtimes and low memory usage. Unlike general purpose a
cceleration frameworks such as XLA, our library turns generic Python code into b
inaries whose performances are competitive with state-of-the-art geometric libra
ries - such as FAISS for nearest neighbor search - with the added benefit of fle
xibility. We perform an extensive evaluation on a broad class of problems: Gauss
ian modelling, K-nearest neighbors search, geometric deep learning, non-Euclidea
n embeddings and optimal transport theory. In practice, for geometric problems t
hat involve 1k to 1M samples in dimension 1 to 100, our library speeds up baseli
ne GPU implementations by up to two orders of magnitude.
**********

MESA: Boost Ensemble Imbalanced Learning with MEta-SAmpler
Zhining Liu, Pengfei Wei, Jing Jiang, Wei Cao, Jiang Bian, Yi Chang
Imbalanced learning (IL), i.e., learning unbiased models from class-imbalanced d
ata, is a challenging problem. Typical IL methods including resampling and rewei
ghting were designed based on some heuristic assumptions. They often suffer from
 unstable performance, poor applicability, and high computational cost in comple
x tasks where their assumptions do not hold. In this paper, we introduce a novel
 ensemble IL framework named MESA. It adaptively resamples the training set in i
terations to get multiple classifiers and forms a cascade ensemble model. MESA d
irectly learns the sampling strategy from data to optimize the final metric beyo
nd following random heuristics. Moreover, unlike prevailing meta-learning-based
IL solutions, we decouple the model-training and meta-training in MESA by indepe
ndently train the meta-sampler over task-agnostic meta-data. This makes MESA gen
erally applicable to most of the existing learning models and the meta-sampler c
an be efficiently applied to new tasks. Extensive experiments on both synthetic
and real-world tasks demonstrate the effectiveness, robustness, and transferabil
ity of MESA. Our code is available at https://github.com/ZhiningLiu1998/mesa.
**********

CoinPress: Practical Private Mean and Covariance Estimation
Sourav Biswas, Yihe Dong, Gautam Kamath, Jonathan Ullman
We present simple differentially private estimators for the parameters of multiv
ariate sub-Gaussian data that are accurate at small sample sizes.  We demonstrat
e the effectiveness of our algorithms both theoretically and empirically using s
ynthetic and real-world datasets---showing that their asymptotic error rates mat
ch the state-of-the-art theoretical bounds, and that they concretely outperform
all previous methods. Specifically, previous estimators either have weak empiric
al accuracy at small sample sizes, perform poorly for multivariate data, or requ
ire the user to provide strong a priori estimates for the parameters.
**********

Planning with General Objective Functions: Going Beyond Total Rewards
Ruosong Wang, Peilin Zhong, Simon S. Du, Russ R. Salakhutdinov, Lin 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.
**********

Scattering GCN: Overcoming Oversmoothness in Graph Convolutional Networks
Yimeng Min, Frederik Wenkel, Guy Wolf
Graph convolutional networks (GCNs) have shown promising results in processing g
raph data by extracting structure-aware features. This gave rise to extensive wo
rk in geometric deep learning, focusing on designing network architectures that
ensure neuron activations conform to regularity patterns within the input graph.
 However, in most cases the graph structure is only accounted for by considering
 the similarity of activations between adjacent nodes, which limits the capabili
ties of such methods to discriminate between nodes in a graph. Here, we propose
 to augment conventional GCNs with geometric scattering transforms and residual
convolutions. The former enables band-pass filtering of graph signals, thus alle
viating the so-called oversmoothing often encountered in GCNs, while the latter
is introduced to clear the resulting features of high-frequency noise. We establ
ish the advantages of the presented Scattering GCN with both theoretical results

establishing the complementary benefits of scattering and GCN features, as well as experimental results showing the benefits of our method compared to leading graph neural networks for semi-supervised node classification, including the rec ently proposed GAT network that typically alleviates oversmoothing using graph a ttention mechanisms.
**********

KFC: A Scalable Approximation Algorithm for $k$–center Fair Clustering
Elfarouk Harb, Ho Shan Lam
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Leveraging Predictions in Smoothed Online Convex Optimization via Gradient-base d Algorithms
Yingying Li, Na 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.
**********

Learning the Linear Quadratic Regulator from Nonlinear Observations
Zakaria Mhammedi, Dylan J. Foster, Max Simchowitz, Dipendra Misra, Wen Sun, Aksh ay Krishnamurthy, Alexander Rakhlin, John Langford
We introduce a new problem setting for continuous control called the LQR with Ri ch Observations, or RichLQR. In our setting, the environment is summarized by a low-dimensional continuous latent state with linear dynamics and quadratic costs , but the agent operates on high-dimensional, nonlinear observations such as ima ges from a camera. To enable sample-efficient learning, we assume that the learn er has access to a class of decoder functions (e.g., neural networks) that is fl exible enough to capture the mapping from observations to latent states. We intr oduce a new algorithm, RichID, which learns a near-optimal policy for the RichLQ R with sample complexity scaling only with the dimension of the latent state spa ce and the capacity of the decoder function class. RichID is oracle-efficient an d accesses the decoder class only through calls to a least-squares regression or acle. To our knowledge, our results constitute the first provable sample complex ity guarantee for continuous control with an unknown nonlinearity in the system model.
**********

Reconciling Modern Deep Learning with Traditional Optimization Analyses: The Int rinsic Learning Rate
Zhiyuan Li, Kaifeng Lyu, Sanjeev 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-auth ors prior to requesting a name change in the electronic proceedings.
**********

Scalable Graph Neural Networks via Bidirectional Propagation
Ming Chen, Zhewei Wei, Bolin Ding, Yaliang Li, Ye Yuan, Xiaoyong Du, Ji-Rong Wen
Graph Neural Networks (GNN) are an emerging field for learning on non-Euclidean data. Recently, there has been increased interest in designing GNN that scales t o large graphs. Most existing methods use "graph sampling" or "layer-wise sampli ng" techniques to reduce training time; However, these methods still suffer from degrading performance and scalability problems when applying to graphs with bil lions of edges. In this paper, we present GBP, a scalable GNN that utilizes a lo calized bidirectional propagation process from both the feature vector and the t raining/testing nodes. Theoretical analysis shows that GBP is the first method t hat achieves sub-linear time complexity for both the precomputation and the trai ning phases. An extensive empirical study demonstrates that GBP achieves state-o f-the-art performance with significantly less training/testing time. Most notabl

y, GBP is able to deliver superior performance on a graph with over 60 million n
odes and 1.8 billion edges in less than 2,000 seconds on a single machine.
**********

Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Le
arning

Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, Jinwoo Shin

While semi-supervised learning (SSL) has proven to be a promising way for levera
ging unlabeled data when labeled data is scarce, the existing SSL algorithms typ
ically assume that training class distributions are balanced. However, these SSL
 algorithms trained under imbalanced class distributions can severely suffer whe
n generalizing to a balanced testing criterion, since they utilize biased pseudo
-labels of unlabeled data toward majority classes. To alleviate this issue, we f
ormulate a convex optimization problem to softly refine the pseudo-labels genera
ted from the biased model, and develop a simple algorithm, named Distribution Al
igning Refinery of Pseudo-label (DARP) that solves it provably and efficiently.
Under various class imbalanced semi-supervised scenarios, we demonstrate the eff
ectiveness of DARP and its compatibility with state-of-the-art SSL schemes.
**********

Assisted Learning: A Framework for Multi-Organization Learning

Xun Xian, Xinran Wang, Jie Ding, Reza Ghanadan

In an increasing number of AI scenarios, collaborations among different organiza
tions or agents (e.g., human and robots, mobile units) are often essential to ac
complish an organization-specific mission. However, to avoid leaking useful and
possibly proprietary information, organizations typically enforce stringent secu
rity constraints on sharing modeling algorithms and data, which significantly li
mits collaborations. In this work, we introduce the Assisted Learning framework
for organizations to assist each other in supervised learning tasks without reve
aling any organization's algorithm, data, or even task. An organization seeks as
sistance by broadcasting task-specific but nonsensitive statistics and incorpora
ting others' feedback in one or more iterations to eventually improve its predic
tive performance. Theoretical and experimental studies, including real-world med
ical benchmarks, show that Assisted Learning can often achieve near-oracle learn
ing performance as if data and training processes were centralized.
**********

The Strong Screening Rule for SLOPE

Johan Larsson, Malgorzata Bogdan, Jonas Wallin

Extracting relevant features from data sets where the number of observations (n)
 is much smaller then the number of predictors (p) is a major challenge in moder
n statistics. Sorted L-One Penalized Estimation (SLOPE)—a generalization of the
lasso—is a promising method within this setting. Current numerical procedures fo
r SLOPE, however, lack the efficiency that respective tools for the lasso enjoy,
 particularly in the context of estimating a complete regularization path. A key
 component in the efficiency of the lasso is predictor screening rules: rules th
at allow  predictors to be discarded before estimating the model. This is the fi
rst paper to establish such a rule for SLOPE. We develop a screening rule for SL
OPE by examining its subdifferential and show that this rule is a generalization
 of the strong rule for the lasso. Our rule is heuristic, which means that it ma
y discard predictors erroneously. In our paper, however, we show that such situa
tions are rare and easily safeguarded against by a simple check of the optimalit
y conditions. Our numerical experiments show that the rule performs well in prac
tice, leading to improvements by orders of magnitude for data in the (p >> n) do
main, as well as incurring no additional computational overhead when (n > p).
**********

STLnet: Signal Temporal Logic Enforced Multivariate Recurrent Neural Networks

Meiyi Ma, Ji Gao, Lu Feng, John Stankovic

Recurrent Neural Networks (RNNs) have made great achievements for sequential pre
diction tasks. In practice, the target sequence often follows certain model prop
erties or patterns (e.g., reasonable ranges, consecutive changes, resource const
raint, temporal correlations between multiple variables, existence, unusual case
s, etc.). However, RNNs cannot guarantee their learned distributions satisfy the

se model properties. It is even more challenging for predicting large-scale and complex Cyber-Physical Systems. Failure to produce outcomes that meet these model properties will result in inaccurate and even meaningless results. In this paper, we develop a new temporal logic-based learning framework, STLnet, which guides the RNN learning process with auxiliary knowledge of model properties, and produces a more robust model for improved future predictions. Our framework can be applied to general sequential deep learning models, and trained in an end-to-end manner with back-propagation. We evaluate the performance of STLnet using large-scale real-world city data. The experimental results show STLnet not only improves the accuracy of predictions, but importantly also guarantees the satisfaction of model properties and increases the robustness of RNNs.
**********

Election Coding for Distributed Learning: Protecting SignSGD against Byzantine Attacks

Jy-yong Sohn, Dong-Jun Han, Beongjun Choi, Jaekyun Moon

Current distributed learning systems suffer from serious performance degradation under Byzantine attacks. This paper proposes Election Coding, a coding-theoretic framework to guarantee Byzantine-robustness for distributed learning algorithms based on signed stochastic gradient descent (SignSGD) that minimizes the worker-master communication load. The suggested framework explores new information-theoretic limits of finding the majority opinion when some workers could be attacked by adversary, and paves the road to implement robust and communication-efficient distributed learning algorithms. Under this framework, we construct two types of codes, random Bernoulli codes and deterministic algebraic codes, that tolerate Byzantine attacks with a controlled amount of computational redundancy and guarantee convergence in general non-convex scenarios.
For the Bernoulli codes, we provide an upper bound on the error probability in estimating the signs of the true gradients, which gives useful insights into code design for Byzantine tolerance. The proposed deterministic codes are proven to perfectly tolerate arbitrary Byzantine attacks. Experiments on real datasets confirm that the suggested codes provide substantial improvement in Byzantine tolerance of distributed learning systems employing SignSGD.
**********

Reducing Adversarially Robust Learning to Non-Robust PAC Learning

Omar Montasser, Steve Hanneke, Nati Srebro

We study the problem of reducing adversarially robust learning to standard PAC learning, i.e. the complexity of learning adversarially robust predictors using access to only a black-box non-robust learner. We give a reduction that can robustly learn any hypothesis class C using any non-robust learner A for C. The number of calls to A depends logarithmically on the number of allowed adversarial perturbations per example, and we give a lower bound showing this is unavoidable.
**********

Top-k Training of GANs: Improving GAN Performance by Throwing Away Bad Samples

Samarth Sinha, Zhengli Zhao, Anirudh Goyal ALIAS PARTH GOYAL, Colin A. Raffel, Augustus Odena

We introduce a simple (one line of code) modification to the Generative Adversarial Network (GAN) training algorithm that materially improves results with no increase in computational cost. When updating the generator parameters, we simply zero out the gradient contributions from the elements of the batch that the critic scores as least realistic'. Through experiments on many different GAN variants, we show that thistop-k update' procedure is a generally applicable improvement. In order to understand the nature of the improvement, we conduct extensive analysis on a simple mixture-of-Gaussians dataset and discover several interesting phenomena. Among these is that, when gradient updates are computed using the worst-scoring batch elements, samples can actually be pushed further away from the their nearest mode. We also apply our method to state-of-the-art GAN models including BigGAN and improve state-of-the-art FID for conditional generation on CIFAR-10 from 9.21 to 8.57.
**********

Black-Box Optimization with Local Generative Surrogates

Sergey Shirobokov, Vladislav Belavin, Michael Kagan, Andrei Ustyuzhanin, Atilim Gunes Baydin

We propose a novel method for gradient-based optimization of black-box simulators using differentiable local surrogate models. In fields such as physics and engineering, many processes are modeled with non-differentiable simulators with intractable likelihoods. Optimization of these forward models is particularly challenging, especially when the simulator is stochastic. To address such cases, we introduce the use of deep generative models to iteratively approximate the simulator in local neighborhoods of the parameter space. We demonstrate that these local surrogates can be used to approximate the gradient of the simulator, and thus enable gradient-based optimization of simulator parameters. In cases where the dependence of the simulator on the parameter space is constrained to a low dimensional submanifold, we observe that our method attains minima faster than baseline methods, including Bayesian optimization, numerical optimization and approaches using score function gradient estimators.

**********

Efficient Generation of Structured Objects with Constrained Adversarial Networks

Luca Di Liello, Pierfrancesco Ardino, Jacopo Gobbi, Paolo Morettin, Stefano Teso, Andrea Passerini

Generative Adversarial Networks (GANs) struggle to generate structured objects like molecules and game maps. The issue is that structured objects must satisfy hard requirements (e.g., molecules must be chemically valid) that are difficult to acquire from examples alone. As a remedy, we propose Constrained Adversarial Networks (CANs), an extension of GANs in which the constraints are embedded into the model during training. This is achieved by penalizing the generator proportionally to the mass it allocates to invalid structures. In contrast to other generative models, CANs support efficient inference of valid structures (with high probability) and allows to turn on and off the learned constraints at inference time. CANs handle arbitrary logical constraints and leverage knowledge compilation techniques to efficiently evaluate the disagreement between the model and the constraints. Our setup is further extended to hybrid logical-neural constraints for capturing very complex constraints, like graph reachability. An extensive empirical analysis shows that CANs efficiently generate valid structures that are both high-quality and novel.

**********

Hard Example Generation by Texture Synthesis for Cross-domain Shape Similarity Learning

Huan Fu, Shunming Li, Rongfei Jia, Mingming Gong, Binqiang Zhao, Dacheng Tao

Image-based 3D shape retrieval (IBSR) aims to find the corresponding 3D shape of a given 2D image from a large 3D shape database. The common routine is to map 2D images and 3D shapes into an embedding space and define (or learn) a shape similarity measure. While metric learning with some adaptation techniques seems to be a natural solution to shape similarity learning, the performance is often unsatisfactory for fine-grained shape retrieval. In the paper, we identify the source of the poor performance and propose a practical solution to this problem. We find that the shape difference between a negative pair is entangled with the texture gap, making metric learning ineffective in pushing away negative pairs. To tackle this issue, we develop a geometry-focused multi-view metric learning framework empowered by texture synthesis. The synthesis of textures for 3D shape models creates hard triplets, which suppress the adverse effects of rich texture in 2D images, thereby push the network to focus more on discovering geometric characteristics. Our approach shows state-of-the-art performance on a recently released large-scale 3D-FUTURE [1] repository, as well as three widely studied benchmarks, including Pix3D [2], Stanford Cars [3], and Comp Cars [4]. Codes will be made publicly available at: https://github.com/3D-FRONT-FUTURE/IBSR-texture.

**********

Recovery of sparse linear classifiers from mixture of responses

Venkata Gandikota, Arya Mazumdar, Soumyabrata Pal

In the problem of learning a mixture of linear classifiers, the aim is to learn a collection of hyperplanes from a sequence of binary responses. Each response i

s a result of querying with a vector and indicates the side of a randomly chosen hyperplane from the collection the query vector belong to. This model is quite rich while dealing with heterogeneous data with categorical labels and has only been studied in some special settings. We look at a hitherto unstudied problem of query complexity upper bound of recovering all the hyperplanes, especially for the case when the hyperplanes are sparse. This setting is a natural generalization of the extreme quantization problem known as 1-bit compressed sensing. Suppose we have a set of l unknown k-sparse vectors. We can query the set with another vector a, to obtain the sign of the inner product of a and a randomly chosen vector from the l-set. How many queries are sufficient to identify all the l unknown vectors? This question is significantly more challenging than both the basic 1-bit compressed sensing problem (i.e., l = 1 case) and the analogous regression problem (where the value instead of the sign is provided). We provide rigorous query complexity results (with efficient algorithms) for this problem.

**********

Efficient Distance Approximation for Structured High-Dimensional Distributions via Learning

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S Meel, N. V. Vinodchandran

We design efficient distance approximation algorithms for several classes of well-studied structured high-dimensional distributions. Specifically, we present algorithms for the following problems (where dTV is the total variation distance):

**********

A Single Recipe for Online Submodular Maximization with Adversarial or Stochastic Constraints

Omid Sadeghi, Prasanna Raut, 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-authors prior to requesting a name change in the electronic proceedings.

**********

Learning Sparse Prototypes for Text Generation

Junxian He, Taylor Berg-Kirkpatrick, Graham Neubig

Prototype-driven text generation uses non-parametric models that first choose from a library of sentence "prototypes" and then modify the prototype to generate the output text. While effective, these methods are inefficient at test time as a result of needing to store and index the entire training corpus. Further, existing methods often require heuristics to identify which prototypes to reference at training time. In this paper, we propose a novel generative model that automatically learns a sparse prototype support set that, nonetheless, achieves strong language modeling performance. This is achieved by (1) imposing a sparsity-inducing prior on the prototype selection distribution, and (2) utilizing amortized variational inference to learn a prototype retrieval function. In experiments, our model outperforms previous prototype-driven language models while achieving up to a 1000x memory reduction, as well as a 1000x speed-up at test time. More interestingly, we show that the learned prototypes are able to capture semantics and syntax at different granularity as we vary the sparsity of prototype selection, and that certain sentence attributes can be controlled by specifying the prototype for generation.

**********

Implicit Rank-Minimizing Autoencoder

Li Jing, Jure Zbontar, yann lecun

An important component of autoencoder methods is the method by which the information capacity of the latent representation is minimized or limited. In this work, the rank of the covariance matrix of the codes is implicitly minimized by relying on the fact that gradient descent learning in multi-layer linear networks leads to minimum-rank solutions. By inserting a number of extra linear layers between the encoder and the decoder, the system spontaneously learns representations with a low effective dimension. The model, dubbed Implicit Rank-Minimizing Autoencoder (IRMAE), is simple, deterministic, and learns continuous latent space. We demonstrate the validity of the method on several image generation and represe

ntation learning tasks.
**********

Storage Efficient and Dynamic Flexible Runtime Channel Pruning via Deep Reinforcement Learning
Jianda Chen, Shangyu Chen, Sinno Jialin Pan
In this paper, we propose a deep reinforcement learning (DRL) based framework to efficiently perform runtime channel pruning on convolutional neural networks (CNNs). Our DRL-based framework aims to learn a pruning strategy to determine how many and which channels to be pruned in each convolutional layer, depending on each individual input instance at runtime. Unlike existing runtime pruning methods which require to store all channels parameters for inference, our framework can reduce parameters storage consumption by introducing a static pruning component. Comparison experimental results with existing runtime and static pruning methods on state-of-the-art CNNs demonstrate that our proposed framework is able to provide a tradeoff between dynamic flexibility and storage efficiency in runtime channel pruning.
**********

Task-Oriented Feature Distillation
Linfeng Zhang, Yukang Shi, Zuoqiang Shi, Kaisheng Ma, Chenglong Bao
Feature distillation, a primary method in knowledge distillation, always leads to significant accuracy improvements. Most existing methods distill features in the teacher network through a manually designed transformation. In this paper, we propose a novel distillation method named task-oriented feature distillation (TOFD) where the transformation is convolutional layers that are trained in a data-driven manner by task loss. As a result, the task-oriented information in the features can be captured and distilled to students. Moreover, an orthogonal loss is applied to the feature resizing layer in TOFD to improve the performance of knowledge distillation. Experiments show that TOFD outperforms other distillation methods by a large margin on both image classification and 3D classification tasks. Codes have been released in Github.
**********

Entropic Causal Inference: Identifiability and Finite Sample Results
Spencer Compton, Murat Kocaoglu, Kristjan Greenewald, Dmitriy Katz
Entropic causal inference is a framework for inferring the causal direction between two categorical variables from observational data. The central assumption is that the amount of unobserved randomness in the system is not too large. This unobserved randomness is measured by the entropy of the exogenous variable in the underlying structural causal model, which governs the causal relation between the observed variables. Kocaoglu et al. conjectured that the causal direction is identifiable when the entropy of the exogenous variable is not too large. In this paper, we prove a variant of their conjecture. Namely, we show that for almost all causal models where the exogenous variable has entropy that does not scale with the number of states of the observed variables, the causal direction is identifiable from observational data. We also consider the minimum entropy coupling-based algorithmic approach presented by Kocaoglu et al., and for the first time demonstrate algorithmic identifiability guarantees using a finite number of samples. We conduct extensive experiments to evaluate the robustness of the method to relaxing some of the assumptions in our theory and demonstrate that both the constant-entropy exogenous variable and the no latent confounder assumptions can be relaxed in practice. We also empirically characterize the number of observational samples needed for causal identification. Finally, we apply the algorithm on Tuebingen cause-effect pairs dataset.
**********

Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement
Ben Eysenbach, XINYANG GENG, Sergey Levine, Russ R. Salakhutdinov
Multi-task reinforcement learning (RL) aims to simultaneously learn policies for solving many tasks. Several prior works have found that relabeling past experience with different reward functions can improve sample efficiency. Relabeling methods typically pose the question: if, in hindsight, we assume that our experience was optimal for some task, for what task was it optimal? Inverse RL answers t

his question. In this paper we show that inverse RL is a principled mechanism fo r reusing experience across tasks. We use this idea to generalize goal-relabelin g techniques from prior work to arbitrary types of reward functions. Our experim ents confirm that relabeling data using inverse RL outperforms prior relabeling methods on goal-reaching tasks, and accelerates learning on more general multi-t ask settings where prior methods are not applicable, such as domains with discre te sets of rewards and those with linear reward functions.
**********

Variance-Reduced Off-Policy TDC Learning: Non-Asymptotic Convergence Analysis
Shaocong Ma, Yi Zhou, Shaofeng 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-auth ors prior to requesting a name change in the electronic proceedings.
**********

AdaTune: Adaptive Tensor Program Compilation Made Efficient
Menghao Li, Minjia Zhang, Chi Wang, Mingqin Li
Deep learning models are computationally intense, and implementations often have to be highly optimized by experts or hardware vendors to be usable in practice. The DL compiler, together with Learning to Compile have proven to be a powerful technique for optimizing tensor programs. However, a limitation of this approac h is that it still suffers from unbearably long overall optimization time.
**********

When Do Neural Networks Outperform Kernel Methods?
Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, Andrea Montanari
For a certain scaling of the initialization of stochastic gradient descent (SGD) , wide neural networks (NN) have been shown to be well approximated by reproduci ng kernel Hilbert space (RKHS) methods. Recent empirical work showed that, for some classification tasks, RKHS methods can replace NNs without a large loss in performance. On the other hand, two-layers NNs are known to encode richer smooth ness classes than RKHS and we know of special examples for which SGD-trained NN provably outperform RKHS. This is true even in the wide network limit, for a dif ferent scaling of the initialization.
**********

STEER : Simple Temporal Regularization For Neural ODE
Arnab Ghosh, Harkirat Behl, Emilien Dupont, Philip Torr, Vinay Namboodiri
Training Neural Ordinary Differential Equations (ODEs) is often computationally expensive. Indeed, computing the forward pass of such models involves solving an ODE which can become arbitrarily complex during training. Recent works have sho wn that regularizing the dynamics of the ODE can partially alleviate this. In th is paper we propose a new regularization technique: randomly sampling the end ti me of the ODE during training. The proposed regularization is simple to implemen t, has negligible overhead and is effective across a wide variety of tasks. Furt her, the technique is orthogonal to several other methods proposed to regularize the dynamics of ODEs and as such can be used in conjunction with them. We show through experiments on normalizing flows, time series models and image recogniti on that the proposed regularization can significantly decrease training time and even improve performance over baseline models.
**********

A Variational Approach for Learning from Positive and Unlabeled Data
Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, Hao Wu
Learning binary classi■ers only from positive and unlabeled (PU) data is an impo rtant and challenging task in many real-world applications, including web text c lassi■cation, disease gene identi■cation and fraud detection, where negative sam ples are dif■cult to verify experimentally. Most recent PU learning methods are developed based on the misclassi■cation risk of the supervised learning type, an d they may suffer from inaccurate estimates of class prior probabilities. In thi s paper, we introduce a variational principle for PU learning that allows us to quantitatively evaluate the modeling error of the Bayesian classi■er directly fr om given data. This leads to a loss function which can be ef■ciently calculated

without involving class prior estimation or any other intermediate estimation pr
oblems, and the variational learning method can then be employed to optimize the
 classi■er under general conditions. We illustrate the effectiveness of the prop
osed variational method on a number of benchmark examples.
**********
Efficient Clustering Based On A Unified View Of $K$-means And Ratio-cut
Shenfei Pei, Feiping Nie, Rong Wang, Xuelong 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.
**********
Recurrent Switching Dynamical Systems Models for Multiple Interacting Neural Pop
ulations
Joshua Glaser, Matthew Whiteway, John P. Cunningham, Liam Paninski, Scott Linder
man
Modern recording techniques can generate large-scale measurements of multiple ne
ural populations over extended time periods. However, it remains a challenge to
model non-stationary interactions between high-dimensional populations of neuron
s. To tackle this challenge, we develop recurrent switching linear dynamical sys
tems models for multiple populations. Here, each high-dimensional neural populat
ion is represented by a unique set of latent variables, which evolve dynamically
 in time. Populations interact with each other through this low-dimensional spac
e. We allow the nature of these interactions to change over time by using a disc
rete set of dynamical states. Additionally, we parameterize these discrete state
 transition rules to capture which neural populations are responsible for switch
ing between interaction states. To fit the model, we use variational expectation
-maximization with a structured mean-field approximation. After validating the m
odel on simulations, we apply it to two different neural datasets: spiking activ
ity from motor areas in a non-human primate, and calcium imaging from neurons in
 the nematode \textit{C. elegans}. In both datasets, the model reveals behaviora
lly-relevant discrete states with unique inter-population interactions and diffe
rent populations that predict transitioning between these states.
**********
Coresets via Bilevel Optimization for Continual Learning and Streaming
Zalán Borsos, Mojmir Mutny, Andreas Krause
Coresets are small data summaries that are sufficient for model training. They c
an be maintained online, enabling efficient handling of large data streams under
 resource constraints. However, existing constructions are limited to simple mod
els such as k-means and logistic regression. In this work, we propose a novel co
reset construction via cardinality-constrained bilevel optimization. We show how
 our framework can efficiently generate coresets for deep neural networks, and d
emonstrate its empirical benefits in continual learning and in streaming setting
s.
**********
Generalized Independent Noise Condition for Estimating Latent Variable Causal Gr
aphs
Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, Kun 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.
**********
Understanding and Exploring the Network with Stochastic Architectures
Zhijie Deng, Yinpeng Dong, Shifeng Zhang, Jun Zhu
There is an emerging trend to train a network with stochastic architectures to e
nable various architectures to be plugged and played during inference. However,
the existing investigation is highly entangled with neural architecture search (
NAS), limiting its widespread use across scenarios. In this work, we decouple th
e training of a network with stochastic architectures (NSA) from NAS and provide

a first systematical investigation on it as a stand-alone problem. We first unc over the characteristics of NSA in various aspects ranging from training stabili ty, convergence, predictive behaviour, to generalization capacity to unseen arch itectures. We identify various issues of the vanilla NSA, such as training/test disparity and function mode collapse, and further propose the solutions to these issues with theoretical and empirical insights. We believe that these results c ould also serve as good heuristics for NAS. Given these understandings, we furth er apply the NSA with our improvements into diverse scenarios to fully exploit i ts promise of inference-time architecture stochasticity, including model ensembl e, uncertainty estimation and semi-supervised learning. Remarkable performance ( e.g., 2.75% error rate and 0.0032 expected calibration error on CIFAR-10) valida te the effectiveness of such a model, providing new perspectives of exploring th e potential of the network with stochastic architectures, beyond NAS.

**********

All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation

jean barbier, Nicolas Macris, Cynthia Rush

We determine statistical and computational limits for estimation of a rank-one m atrix (the spike) corrupted by an additive gaussian noise matrix, in a sparse li mit, where the underlying hidden vector (that constructs the rank-one matrix) ha s a number of non-zero components that scales sub-linearly with the total dimens ion of the vector, and the signal-to-noise ratio tends to infinity at an appropr iate speed. We prove explicit low-dimensional variational formulas for the asymp totic mutual information between the spike and the observed noisy matrix and ana lyze the approximate message passing algorithm in the sparse regime. For Bernoul li and Bernoulli-Rademacher distributed vectors, and when the sparsity and signa l strength satisfy an appropriate scaling relation, we find all-or-nothing phase transitions for the asymptotic minimum and algorithmic mean-square-errors. Thes e jump from their maximum possible value to zero, at well defined signal-to-nois e thresholds whose asymptotic values we determine exactly. In the asymptotic reg ime the statistical-to-algorithmic gap diverges indicating that sparse recovery is hard for approximate message passing.

**********

Deep Evidential Regression

Alexander Amini, Wilko Schwarting, Ava Soleimany, Daniela Rus

Deterministic neural networks (NNs) are increasingly being deployed in safety cr itical domains, where calibrated, robust, and efficient measures of uncertainty are crucial. In this paper, we propose a novel method for training non-Bayesian NNs to estimate a continuous target as well as its associated evidence in order to learn both aleatoric and epistemic uncertainty. We accomplish this by placin g evidential priors over the original Gaussian likelihood function and training the NN to infer the hyperparameters of the evidential distribution. We additiona lly impose priors during training such that the model is regularized when its pr edicted evidence is not aligned with the correct output. Our method does not rel y on sampling during inference or on out-of-distribution (OOD) examples for trai ning, thus enabling efficient and scalable uncertainty learning. We demonstrate learning well-calibrated measures of uncertainty on various benchmarks, scaling to complex computer vision tasks, as well as robustness to adversarial and OOD t est samples.

**********

Analytical Probability Distributions and Exact Expectation-Maximization for Deep Generative Networks

Randall Balestriero, Sebastien PARIS, Richard Baraniuk

Deep Generative Networks (DGNs) with probabilistic modeling of their output and latent space are currently trained via Variational Autoencoders (VAEs).

In the absence of a known analytical form for the posterior and likelihood expectation, VAEs resort to approximations, including (Amortized) Variational Inference (AVI) and Monte-Carlo

sampling.
We exploit the Continuous Piecewise Affine
property
of modern DGNs to derive their posterior and marginal
distributions as well as the latter's first two moments.
These findings enable us
to derive an analytical Expectation-Maximization (EM) algorithm for gradient-free DGN learning.
We demonstrate empirically that EM training of DGNs produces greater
likelihood than VAE training.
Our new framework will guide the design of new VAE AVI that better approximates the true posterior and open new avenues to apply standard statistical tools for model comparison, anomaly detection, and missing data imputation.
**********

Bayesian Pseudocoresets
Dionysis Manousakas, Zuheng Xu, Cecilia Mascolo, Trevor Campbell
Standard Bayesian inference algorithms are prohibitively expensive in the regime of modern large-scale data. Recent work has found that a small, weighted subset of data (a coreset) may be used in place of the full dataset during inference, taking advantage of data redundancy to reduce computational cost. However, this approach has limitations in the increasingly common setting of sensitive, high-dimensional data. Indeed, we prove that there are situations in which the Kullback-Leibler divergence between the optimal coreset and the true posterior grows with data dimension; and as coresets include a subset of the original data, they cannot be constructed in a manner that preserves individual privacy. We address both of these issues with a single unified solution, Bayesian pseudocoresets --- a small weighted collection of synthetic "pseudodata"---along with a variational optimization method to select both pseudodata and weights. The use of pseudodata (as opposed to the original datapoints) enables both the summarization of high-dimensional data and the differentially private summarization of sensitive data. Real and synthetic experiments on high-dimensional data demonstrate that Bayesian pseudocoresets achieve significant improvements in posterior approximation error compared to traditional coresets, and that pseudocoresets provide privacy without a significant loss in approximation quality.
**********

See, Hear, Explore: Curiosity via Audio-Visual Association
Victoria Dean, Shubham Tulsiani, Abhinav Gupta
Exploration is one of the core challenges in reinforcement learning. A common formulation of curiosity-driven exploration uses the difference between the real future and the future predicted by a learned model. However, predicting the future is an inherently difficult task which can be ill-posed in the face of stochasticity. In this paper, we introduce an alternative form of curiosity that rewards novel associations between different senses. Our approach exploits multiple modalities to provide a stronger signal for more efficient exploration. Our method is inspired by the fact that, for humans, both sight and sound play a critical role in exploration. We present results on several Atari environments and Habitat (a photorealistic navigation simulator), showing the benefits of using an audio-visual association model for intrinsically guiding learning agents in the absence of external rewards. For videos and code, see https://vdean.github.io/audio-curiosity.html.
**********

Adversarial Training is a Form of Data-dependent Operator Norm Regularization
Kevin Roth, Yannic Kilcher, Thomas Hofmann
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Biologically Plausible Neural Network for Slow Feature Analysis
David Lipshutz, Charles Windolf, Siavash Golkar, Dmitri Chklovskii

Learning latent features from time series data is an important problem in both m
achine learning and brain function. One approach, called Slow Feature Analysis (
SFA), leverages the slowness of many salient features relative to the rapidly va
rying input signals. Furthermore, when trained on naturalistic stimuli, SFA repr
oduces interesting properties of cells in the primary visual cortex and hippocam
pus, suggesting that the brain uses temporal slowness as a computational princip
le for learning latent features. However, despite the potential relevance of SFA
 for modeling brain function, there is currently no SFA algorithm with a biologi
cally plausible neural network implementation, by which we mean an algorithm ope
rates in the online setting and can be mapped onto a neural network with local s
ynaptic updates. In this work, starting from an SFA objective, we derive an SFA
algorithm, called Bio-SFA, with a biologically plausible neural network implemen
tation. We validate Bio-SFA on naturalistic stimuli.
**********

Learning Feature Sparse Principal Subspace
Lai Tian, Feiping Nie, Rong Wang, Xuelong Li
This paper presents new algorithms to solve the feature-sparsity constrained PCA
 problem (FSPCA), which performs feature selection and PCA simultaneously. Exist
ing optimization methods for FSPCA require data distribution assumptions and are
 lack of global convergence guarantee. Though the general FSPCA problem is NP-ha
rd, we show that, for a low-rank covariance, FSPCA can be solved globally (Algor
ithm 1). Then, we propose another strategy (Algorithm 2) to solve FSPCA for the
general covariance by iteratively building a carefully designed proxy. We prove
(data-dependent) approximation bound and convergence guarantees for the new algo
rithms. For the spectrum of covariance with exponential/Zipf's distribution, we
provide exponential/posynomial approximation bound. Experimental results show th
e promising performance and efficiency of the new algorithms compared with the s
tate-of-the-arts on both synthetic and real-world datasets.
**********

Online Adaptation for Consistent Mesh Reconstruction in the Wild
Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Y
ang, Jan Kautz
This paper presents an algorithm to reconstruct temporally consistent 3D meshes
of deformable object instances from videos in the wild. Without requiring annota
tions of 3D mesh, 2D keypoints, or camera pose for each video frame, we pose vid
eo-based reconstruction as a self-supervised online adaptation problem applied t
o any incoming test video. We first learn a category-specific 3D reconstruction
model from a collection of single-view images of the same category that jointly
predicts the shape, texture, and camera pose of an image. Then, at inference tim
e, we adapt the model to a test video over time using self-supervised regulariza
tion terms that exploit  temporal consistency of an object instance to enforce t
hat all reconstructed meshes share a common texture map, a base shape, as well a
s parts. We demonstrate that our algorithm recovers temporally consistent and re
liable 3D structures from videos of non-rigid objects including those of animals
 captured in the wild -- an extremely challenging task rarely addressed before.
**********

Online learning with dynamics: A minimax perspective
Kush Bhatia, Karthik Sridharan
We consider the problem of online learning with dynamics, where a learner intera
cts with a stateful environment over multiple rounds. In each round of the inter
action, the learner selects a  policy to deploy and incurs a cost that depends o
n both the chosen policy and current state of the world. The state-evolution dyn
amics and the costs are allowed to be time-varying, in a possibly adversarial wa
y. In this setting, we study the problem of minimizing policy regret and provide
 non-constructive upper bounds on the minimax rate for the problem.
**********

Learning to Select Best Forecast Tasks for Clinical Outcome Prediction
Yuan Xue, Nan Du, Anne Mottram, Martin Seneviratne, Andrew M. Dai
The paradigm of pretraining' from a set of relevant auxiliary tasks and thenfine
tuning' on a target task has been successfully applied in many different domains

. However, when the auxiliary tasks are abundant, with complex relationships to the target task, using domain knowledge or searching over all possible pretraining setups are inefficient strategies. To address this challenge, we propose a method to automatically select from a large set of auxiliary tasks which yield a representation most useful to the target task. In particular, we develop an efficient algorithm that uses automatic auxiliary task selection within a nested-loop meta-learning process. We have applied this algorithm to the task of clinical outcome predictions in electronic medical records, learning from a large number of self-supervised tasks related to forecasting patient trajectories. Experiments on a real clinical dataset demonstrate the superior predictive performance of our method compared to direct supervised learning, naive pretraining and multitask learning, in particular in low-data scenarios when the primary task has very few examples. With detailed ablation analysis, we further show that the selection rules are interpretable and able to generalize to unseen target tasks with new data.

**********

Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping

Eduard Gorbunov, Marina Danilova, Alexander Gasnikov

In this paper, we propose a new accelerated stochastic first-order method called clipped-SSTM for smooth convex stochastic optimization with heavy-tailed distributed noise in stochastic gradients and derive the first high-probability complexity bounds for this method closing the gap in the theory of stochastic optimization with heavy-tailed noise. Our method is based on a special variant of accelerated Stochastic Gradient Descent (SGD) and clipping of stochastic gradients. We extend our method to the strongly convex case and prove new complexity bounds that outperform state-of-the-art results in this case. Finally, we extend our proof technique and derive the first non-trivial high-probability complexity bounds for SGD with clipping without light-tails assumption on the noise.

**********

Adaptive Experimental Design with Temporal Interference: A Maximum Likelihood Approach

Peter W. Glynn, Ramesh Johari, Mohammad Rasouli

Suppose an online platform wants to compare a treatment and control policy (e.g., two different matching algorithms in a ridesharing system, or two different inventory management algorithms in an online retail site). Standard experimental approaches to this problem are biased (due to temporal interference between the policies), and not sample efficient. We study optimal experimental design for this setting. We view testing the two policies as the problem of estimating the steady state difference in reward between two unknown Markov chains (i.e., policies). We assume estimation of the steady state reward for each chain proceeds via nonparametric maximum likelihood, and search for consistent (i.e., asymptotically unbiased) experimental designs that are efficient (i.e., asymptotically minimum variance). Characterizing such designs is equivalent to a Markov decision problem with a minimum variance objective; such problems generally do not admit tractable solutions. Remarkably, in our setting, using a novel application of classical martingale analysis of Markov chains via Poisson's equation, we characterize efficient designs via a succinct convex optimization problem. We use this characterization to propose a consistent, efficient online experimental design that adaptively samples the two Markov chains.

**********

From Trees to Continuous Embeddings and Back: Hyperbolic Hierarchical Clustering

Ines Chami, Albert Gu, Vaggos Chatziafratis, Christopher Ré

Similarity-based Hierarchical Clustering (HC) is a classical unsupervised machine learning algorithm that has traditionally been solved with heuristic algorithms like Average-Linkage. Recently, Dasgupta reframed HC as a discrete optimization problem by introducing a global cost function measuring the quality of a given tree. In this work, we provide the first continuous relaxation of Dasgupta's discrete optimization problem with provable quality guarantees. The key idea of our method, HypHC, is showing a direct correspondence from discrete trees to conti

nuous representations (via the hyperbolic embeddings of their leaf nodes) and back (via a decoding algorithm that maps leaf embeddings to a dendrogram), allowing us to search the space of discrete binary trees with continuous optimization. Building on analogies between trees and hyperbolic space, we derive a continuous analogue for the notion of lowest common ancestor, which leads to a continuous relaxation of Dasgupta's discrete objective. We can show that after decoding, the global minimizer of our continuous relaxation yields a discrete tree with a (1+eps)-factor approximation for Dasgupta's optimal tree, where eps can be made arbitrarily small and controls optimization challenges. We experimentally evaluate HypHC on a variety of HC benchmarks and find that even approximate solutions found with gradient descent have superior clustering quality than agglomerative heuristics or other gradient based algorithms. Finally, we highlight the flexibility of HypHC using end-to-end training in a downstream classification task.
**********

The Autoencoding Variational Autoencoder
Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, Pushmeet Kohli
Does a Variational AutoEncoder (VAE) consistently encode typical samples generated from its decoder? This paper shows that the perhaps surprising answer to this question is `No'; a (nominally trained) VAE does not necessarily amortize inference for typical samples that it is capable of generating. We study the implications of this behaviour on the learned representations and also the consequences of fixing it by introducing a notion of self consistency.
Our approach hinges on an alternative construction of the variational approximation distribution to the true posterior of an extended VAE model with a Markov chain alternating between the encoder and the decoder. The method can be used to train a VAE model from scratch or given an already trained VAE, it can be run as a post processing step in an entirely self supervised way without access to the original training data.
Our experimental analysis reveals that encoders trained with our self-consistency approach lead to representations that are robust (insensitive) to perturbations in the input introduced by adversarial attacks.
We provide experimental results on the ColorMnist and CelebA benchmark datasets that quantify the properties of the learned representations and compare the approach with a baseline that is specifically trained for the desired property.
**********

A Fair Classifier Using Kernel Density Estimation
Jaewoong Cho, Gyeongjo Hwang, Changho Suh
As machine learning becomes prevalent in a widening array of sensitive applications such as job hiring and criminal justice, one critical aspect that machine learning classifiers should respect is to ensure fairness: guaranteeing the irrelevancy of a prediction output to sensitive attributes such as gender and race. In this work, we develop a kernel density estimation trick to quantify fairness measures that capture the degree of the irrelevancy. A key feature of our approach is that quantified fairness measures can be expressed as differentiable functions w.r.t. classifier model parameters. This then allows us to enjoy prominent gradient descent to readily solve an interested optimization problem that fully respects fairness constraints. We focus on a binary classification setting and two well-known definitions of group fairness: Demographic Parity (DP) and Equalized Odds (EO). Our experiments both on synthetic and benchmark real datasets demonstrate that our algorithm outperforms prior fair classifiers in accuracy-fairness tradeoff performance both w.r.t. DP and EO.
**********

A Randomized Algorithm to Reduce the Support of Discrete Measures
Francesco Cosentino, Harald Oberhauser, Alessandro Abate
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Distributionally Robust Federated Averaging
Yuyang Deng, Mohammad Mahdi Kamani, Mehrdad Mahdavi
In this paper, we study communication efficient distributed algorithms for distr ibutionally robust federated learning via periodic averaging with adaptive sampl ing. In contrast to standard empirical risk minimization, due to the minimax st ructure of the underlying optimization problem, a key difficulty arises from the fact that the global parameter that controls the mixture of local losses can on ly be updated infrequently on the global stage. To compensate for this, we propo se a Distributionally Robust Federated Averaging (DRFA) algorithm that employs a novel snapshotting scheme to approximate the accumulation of history gradients of the mixing parameter. We analyze the convergence rate of DRFA in both convex-linear and nonconvex-linear settings. We also generalize the proposed idea to ob jectives with regularization on the mixture parameter and propose a proximal va riant, dubbed as DRFA-Prox, with provable convergence rates. We also analyze an alternative optimization method for regularized case in strongly-convex-strongly -concave and non-convex (under PL condition)-strongly-concave settings. To the b est of our knowledge, this paper is the first to solve distributionally robust f ederated learning with reduced communication, and to analyze the efficiency of l ocal descent methods on distributed minimax problems. We give corroborating expe rimental evidence for our theoretical results in federated learning settings.
**********
Sharp uniform convergence bounds through empirical centralization
Cyrus Cousins, Matteo Riondato
We introduce the use of empirical centralization to derive novel practical, prob abilistic, sample-dependent bounds to the Supremum Deviation (SD) of empirical m eans of functions in a family from their expectations.
Our bounds have optimal dependence on the maximum (i.e., wimpy) variance
and the function ranges, and the same dependence on the number of samples as exi sting SD bounds.
To compute the SD bounds in practice, we develop tightly-concentrated
Monte Carlo estimators of the empirical Rademacher average of the empirically-ce ntralized family, and we show novel concentration results for the empirical wimp y variance.
Our experimental evaluation shows that our bounds greatly outperform non-central ized bounds and are extremely practical
even at small sample sizes.
**********
COBE: Contextualized Object Embeddings from Narrated Instructional Video
Gedas Bertasius, Lorenzo Torresani
Many objects in the real world undergo dramatic variations in visual appearance. For example, a tomato may be red or green, sliced or chopped, fresh or fried, l iquid or solid. Training a single detector to accurately recognize tomatoes in a ll these different states is challenging. On the other hand, contextual cues (e. g., the presence of a knife, a cutting board, a strainer or a pan) are often str ongly indicative of how the object appears in the scene. Recognizing such contex tual cues is useful not only to improve the accuracy of object detection or to d etermine the state of the object, but also to understand its functional properti es and to infer ongoing or upcoming human-object interactions. A fully-supervise d approach to recognizing object states and their contexts in the real-world is unfortunately marred by the long-tailed, open-ended distribution of the data, wh ich would effectively require massive amounts of annotations to capture the appe arance of objects in all their different forms. Instead of relying on manually-l abeled data for this task, we propose a new framework for learning Contextualize d OBject Embeddings (COBE) from automatically-transcribed narrations of instruc tional videos. We leverage the semantic and compositional structure of language by training a visual detector to predict a contextualized word embedding of the object and its associated narration. This enables the learning of an object repr esentation where concepts relate according to a semantic language metric. Our ex periments show that our detector learns to predict a rich variety of contextual object information, and that it is highly effective in the settings of few-shot

and zero-shot learning.
**********
Knowledge Transfer in Multi-Task Deep Reinforcement Learning for Continuous Control
Zhiyuan Xu, Kun Wu, Zhengping Che, Jian Tang, Jieping Ye
While Deep Reinforcement Learning (DRL) has emerged as a promising approach to many complex tasks, it remains challenging to train a single DRL agent that is capable of undertaking multiple different continuous control tasks. In this paper, we present a Knowledge Transfer based Multi-task Deep Reinforcement Learning framework (KTM-DRL) for continuous control, which enables a single DRL agent to achieve expert-level performance in multiple different tasks by learning from task-specific teachers. In KTM-DRL, the multi-task agent first leverages an offline knowledge transfer algorithm designed particularly for the actor-critic architecture to quickly learn a control policy from the experience of task-specific teachers, and then it employs an online learning algorithm to further improve itself by learning from new online transition samples under the guidance of those teachers. We perform a comprehensive empirical study with two commonly-used benchmarks in the MuJoCo continuous control task suite. The experimental results well justify the effectiveness of KTM-DRL and its knowledge transfer and online learning algorithms, as well as its superiority over the state-of-the-art by a large margin.
**********
Finite Versus Infinite Neural Networks: an Empirical Study
Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, Jascha Sohl-Dickstein
We perform a careful, thorough, and large scale empirical study of the correspondence between wide neural networks and kernel methods. By doing so, we resolve a variety of open questions related to the study of infinitely wide neural networks. Our experimental results include: kernel methods outperform fully-connected finite-width networks, but underperform convolutional finite width networks; neural network Gaussian process (NNGP) kernels frequently outperform neural tangent (NT) kernels; centered and ensembled finite networks have reduced posterior variance and behave more similarly to infinite networks; weight decay and the use of a large learning rate break the correspondence between finite and infinite networks; the NTK parameterization outperforms the standard parameterization for finite width networks; diagonal regularization of kernels acts similarly to early stopping; floating point precision limits kernel performance beyond a critical dataset size; regularized ZCA whitening improves accuracy; finite network performance depends non-monotonically on width in ways not captured by double descent phenomena; equivariance of CNNs is only beneficial for narrow networks far from the kernel regime. Our experiments additionally motivate an improved layer-wise scaling for weight decay which improves generalization in finite-width networks. Finally, we develop improved best practices for using NNGP and NT kernels for prediction, including a novel ensembling technique. Using these best practices we achieve state-of-the-art results on CIFAR-10 classification for kernels corresponding to each architecture class we consider.
**********
Supermasks in Superposition
Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, Ali Farhadi
We present the Supermasks in Superposition (SupSup) model, capable of sequentially learning thousands of tasks without catastrophic forgetting. Our approach uses a randomly initialized, fixed base network and for each task finds a subnetwork (supermask) that achieves good performance. If task identity is given at test

time, the correct subnetwork can be retrieved with minimal memory usage. If not provided, SupSup can infer the task using gradient-based optimization to find a linear superposition of learned supermasks which minimizes the output entropy. In practice we find that a single gradient step is often sufficient to identify the correct mask, even among 2500 tasks. We also showcase two promising extensions. First, SupSup models can be trained entirely without task identity information, as they may detect when they are uncertain about new data and allocate an additional supermask for the new training distribution. Finally the entire, growing set of supermasks can be stored in a constant-sized reservoir by implicitly storing them as attractors in a fixed-sized Hopfield network.

**********

Nonasymptotic Guarantees for Spiked Matrix Recovery with Generative Priors

Jorio Cocola, Paul Hand, Vlad Voroninski

Many problems in statistics and machine learning require the reconstruction of a rank-one signal matrix from noisy data. Enforcing additional prior information on the rank-one component is often key to guaranteeing good recovery performance. One such prior on the low-rank component is sparsity, giving rise to the sparse principal component analysis problem. Unfortunately, there is strong evidence that this problem suffers from a computational-to-statistical gap, which may be fundamental. In this work, we study an alternative prior where the low-rank component is in the range of a trained generative network. We provide a non-asymptotic analysis with optimal sample complexity, up to logarithmic factors, for rank-one matrix recovery under an expansive-Gaussian network prior. Specifically, we establish a favorable global optimization landscape for a nonlinear least squares objective, provided the number of samples is on the order of the dimensionality of the input to the generative model. This result suggests that generative priors have no computational-to-statistical gap for structured rank-one matrix recovery in the finite data, nonasymptotic regime. We present this analysis in the case of both the Wishart and Wigner spiked matrix models.

**********

Almost Optimal Model-Free Reinforcement Learningvia Reference-Advantage Decomposition

Zihan Zhang, Yuan Zhou, Xiangyang Ji

We study the reinforcement learning problem in the setting of finite-horizon1episodic Markov Decision Processes (MDPs) with S states, A actions, and episode length H. We propose a model-free algorithm UCB-ADVANTAGE and prove that it achieves $\tilde{O}(\sqrt{H^2 SAT})$ regret where T=KH and K is the number of episodes to play. Our regret bound improves upon the results of [Jin et al., 2018] and matches the best known model-based algorithms as well as the information theoretic lower bound up to logarithmic factors. We also show that UCB-ADVANTAGE achieves low local switching cost and applies to concurrent reinforcement learning, improving upon the recent results of [Bai et al., 2019].

**********

Learning to Incentivize Other Learning Agents

Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, Hongyuan Zha

The challenge of developing powerful and general Reinforcement Learning (RL) agents has received increasing attention in recent years. Much of this effort has focused on the single-agent setting, in which an agent maximizes a predefined extrinsic reward function. However, a long-term question inevitably arises: how will such independent agents cooperate when they are continually learning and acting in a shared multi-agent environment? Observing that humans often provide incentives to influence others' behavior, we propose to equip each RL agent in a multi-agent environment with the ability to give rewards directly to other agents, using a learned incentive function. Each agent learns its own incentive function by explicitly accounting for its impact on the learning of recipients and, through them, the impact on its own extrinsic objective. We demonstrate in experiments that such agents significantly outperform standard RL and opponent-shaping agents in challenging general-sum Markov games, often by finding a near-optimal division of labor. Our work points toward more opportunities and challenges along t

he path to ensure the common good in a multi-agent future.
**********
Displacement-Invariant Matching Cost Learning for Accurate Optical Flow Estimation
Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, Hongdong Li
Learning matching costs has been shown to be critical to the success of the state-of-the-art deep stereo matching methods, in which 3D convolutions are applied on a 4D feature volume to learn a 3D cost volume. However, this mechanism has never been employed for the optical flow task. This is mainly due to the significantly increased search dimension in the case of optical flow computation, \ie, a straightforward extension would require dense 4D convolutions in order to process a 5D feature volume, which is computationally prohibitive.
This paper proposes a novel solution that is able to bypass the requirement of building a 5D feature volume while still allowing the network to learn suitable matching costs from data.  Our key innovation is to decouple the connection between 2D displacements and learn the matching costs at each 2D displacement hypothesis independently, \ie, displacement-invariant cost learning. Specifically, we apply the same 2D convolution-based matching net independently on each 2D displacement hypothesis to learn a 4D cost volume. Moreover, we propose a displacement-aware projection layer to scale the learned cost volume, which reconsiders the correlation between different displacement candidates and mitigates the multi-modal problem in the learned cost volume. The cost volume is then projected to optical flow estimation through a 2D soft-argmin layer. Extensive experiments show that our approach achieves state-of-the-art accuracy on various datasets, and outperforms all published optical flow methods on the Sintel benchmark. The code is available at https://github.com/jytime/DICL-Flow.
**********
Distributionally Robust Local Non-parametric Conditional Estimation
Viet Anh Nguyen, Fan Zhang, Jose Blanchet, Erick Delage, Yinyu  Ye
Conditional estimation given specific covariate values (i.e., local conditional estimation or functional estimation) is ubiquitously useful with applications in engineering, social and natural sciences. Existing data-driven non-parametric estimators mostly focus on structured homogeneous data (e.g., weakly independently and stationary data), thus they are sensitive to adversarial noise and may perform poorly under a low sample size. To alleviate these issues, we propose a new distributionally robust estimator that generates non-parametric local estimates by minimizing the worst-case conditional expected loss over all adversarial distributions in a Wasserstein ambiguity set. We show that despite being generally intractable, the local estimator can be efficiently found via convex optimization under broadly applicable settings, and it is robust to the corruption and heterogeneity of the data. Various experiments show the competitive performance of this new class of estimator.
**********
Robust Multi-Object Matching via Iterative Reweighting of the Graph Connection Laplacian
Yunpeng Shi, Shaohan Li, Gilad Lerman
We propose an efficient and robust iterative solution to the multi-object matching problem. We first clarify serious limitations of current methods as well as the inappropriateness of the standard iteratively reweighted least squares procedure. In view of these limitations, we suggest a novel and more reliable iterative reweighting strategy that incorporates information from higher-order neighborhoods by exploiting the graph connection Laplacian. We demonstrate the superior performance of our procedure over state-of-the-art methods using both synthetic and real datasets.
**********
Meta-Gradient Reinforcement Learning with an Objective Discovered Online
Zhongwen Xu, Hado P. van Hasselt, Matteo Hessel, Junhyuk Oh, Satinder Singh, David Silver
Deep reinforcement learning includes a broad family of algorithms that parameterise an internal representation, such as a value function or policy, by a deep ne

ural network. Each algorithm optimises its parameters with respect to an objecti
ve, such as Q-learning or policy gradient, that defines its semantics. In this w
ork, we propose an algorithm based on meta-gradient descent that discovers its o
wn objective, flexibly parameterised by a deep neural network, solely from inter
active experience with its environment. Over time, this allows the agent to lear
n how to learn increasingly effectively. Furthermore, because the objective is d
iscovered online, it can adapt to changes over time. We demonstrate that the alg
orithm discovers how to address several important issues in RL, such as bootstra
pping, non-stationarity, and off-policy learning. On the Atari Learning Environm
ent, the meta-gradient algorithm adapts over time to learn with greater efficien
cy, eventually outperforming the median score of a strong actor-critic baseline.

**********
Learning Strategy-Aware Linear Classifiers
Yiling Chen, Yang Liu, Chara Podimata
We address the question of repeatedly learning linear classifiers against agents
 who are \emph{strategically} trying to \emph{game} the deployed classifiers, an
d we use the \emph{Stackelberg regret} to measure the performance of our algorit
hms. First, we show that Stackelberg and external regret for the problem of stra
tegic classification are \emph{strongly incompatible}: i.e., there exist worst-c
ase scenarios, where \emph{any} sequence of actions providing \emph{sublinear} e
xternal regret might result in \emph{linear} Stackelberg regret and vice versa.
Second, we present a strategy-aware algorithm for minimizing the Stackelberg reg
ret for which we prove nearly matching upper and lower regret bounds. Finally, w
e provide simulations to complement our theoretical analysis. Our results advanc
e the growing literature of learning from revealed preferences, which has so far
 focused on ``smoother'' assumptions from the perspective of the learner and the
 agents respectively.
**********
Upper Confidence Primal-Dual Reinforcement Learning for CMDP with Adversarial Lo
ss
Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********
Calibrating Deep Neural Networks using Focal Loss
Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, P
uneet Dokania
Miscalibration -- a mismatch between a model's confidence and its correctness --
 of Deep Neural Networks (DNNs) makes their predictions hard to rely on. Ideally
, we want networks to be accurate, calibrated and confident. We show that, as op
posed to the standard cross-entropy loss, focal loss (Lin et al., 2017) allows u
s to learn models that are already very well calibrated. When combined with temp
erature scaling, whilst preserving accuracy, it yields state-of-the-art calibrat
ed models. We provide a thorough analysis of the factors causing miscalibration,
 and use the insights we glean from this to justify the empirically excellent pe
rformance of focal loss. To facilitate the use of focal loss in practice, we als
o provide a principled approach to automatically select the hyperparameter invol
ved in the loss function. We perform extensive experiments on a variety of compu
ter vision and NLP datasets, and with a wide variety of network architectures, a
nd show that our approach achieves state-of-the-art calibration without compromi
sing on accuracy in almost all cases. Code is available at https://github.com/to
rrvision/focal_calibration.
**********
Optimizing Mode Connectivity via Neuron Alignment
Norman Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, Rongjie L
ai
The loss landscapes of deep neural networks are not well understood due to their

high nonconvexity. Empirically, the local minima of these loss functions can be connected by a learned curve in model space, along which the loss remains nearly constant; a feature known as mode connectivity. Yet, current curve finding algorithms do not consider the influence of symmetry in the loss surface created by model weight permutations. We propose a more general framework to investigate the effect of symmetry on landscape connectivity by accounting for the weight permutations of the networks being connected. To approximate the optimal permutation, we introduce an inexpensive heuristic referred to as neuron alignment. Neuron alignment promotes similarity between the distribution of intermediate activations of a model along the curve with that of the endpoint models. We provide theoretical analysis establishing the benefit of alignment to mode connectivity based on this simple heuristic. We empirically verify that the permutation given by alignment is locally optimal via a proximal alternating minimization scheme. Empirically, optimizing the weight permutation is critical for efficiently learning a simple, planar, low-loss curve between networks that successfully generalizes. Our alignment method can significantly alleviate the recently identified robust loss barrier on the path connecting two adversarial robust models and find more robust and accurate models on the path.
**********

Information Theoretic Regret Bounds for Online Nonlinear Control
Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, Wen 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-authors prior to requesting a name change in the electronic proceedings.
**********

A kernel test for quasi-independence
Tamara Fernandez, Wenkai Xu, Marc Ditzhaus, Arthur Gretton
We consider settings in which the data of interest correspond to pairs of ordered times, e.g, the birth times of the first and second child, the times at which a new user creates an account and makes the first purchase on a website, and the entry and survival times of patients in a clinical trial. In these settings, the two times are not independent (the second occurs after the first), yet it is still of interest to determine whether there exists significant dependence "beyond" their ordering in time. We refer to this notion as "quasi-(in)dependence." For instance, in a clinical trial, to avoid biased selection, we might wish to verify that recruitment times are quasi-independent of survival times, where dependencies might arise due to seasonal effects. In this paper, we propose a nonparametric statistical test of quasi-independence. Our test considers a potentially infinite space of alternatives, making it suitable for complex data where the nature of the possible quasi-dependence is not known in advance.  Standard parametric approaches are recovered as special cases, such as the classical conditional Kendall's tau, and log-rank tests. The tests apply in the right-censored setting: an essential feature in clinical trials, where patients can withdraw from the study.  We provide an asymptotic analysis of our test-statistic, and demonstrate in experiments that our test obtains better power than existing approaches, while being more computationally efficient.
**********

First Order Constrained Optimization in Policy Space
Yiming Zhang, Quan Vuong, Keith Ross
In reinforcement learning, an agent attempts to learn high-performing behaviors through interacting with the environment, such behaviors are often quantified in the form of a reward function. However some aspects of behavior—such as ones which are deemed unsafe and to be avoided—are best captured through constraints. We propose a novel approach called First Order Constrained Optimization in Policy Space (FOCOPS) which maximizes an agent's overall reward while ensuring the agent satisfies a set of cost constraints. Using data generated from the current policy, FOCOPS first finds the optimal update policy by solving a constrained optimization problem in the nonparameterized policy space. FOCOPS then projects the update policy back into the parametric policy space. Our approach has an approxi

mate upper bound for worst-case constraint violation throughout training and is first-order in nature therefore simple to implement. We provide empirical evidence that our simple approach achieves better performance on a set of constrained robotics locomotive tasks.

**********

Learning Augmented Energy Minimization via Speed Scaling

Etienne Bamas, Andreas Maggiori, Lars Rohwedder, Ola Svensson

As power management has become a primary concern in modern data centers, computing resources are being scaled dynamically to minimize energy consumption. We initiate the study of a variant of the classic online speed scaling problem, in which machine learning predictions about the future can be integrated naturally. Inspired by recent work on learning-augmented online algorithms, we propose an algorithm which incorporates predictions in a black-box manner and outperforms any online algorithm if the accuracy is high, yet maintains provable guarantees if the prediction is very inaccurate.  We provide both theoretical and experimental evidence to support our claims.

**********

Exploiting MMD and Sinkhorn Divergences for Fair and Transferable Representation Learning

Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, Massimiliano Pontil

Developing learning methods which do not discriminate subgroups in the population is a central goal of algorithmic fairness. One way to reach this goal is by modifying the data representation in order to meet certain fairness constraints. In this work we measure fairness according to demographic parity. This requires the probability of the possible model decisions to be independent of the sensitive information.

We argue that the goal of imposing demographic parity can be substantially facilitated within a multitask learning setting. We present a method for learning a shared fair representation across multiple tasks, by means of different new constraints based on MMD and Sinkhorn Divergences. We derive learning bounds establishing that the learned representation transfers well to novel tasks. We present experiments on three real world datasets, showing that the proposed method outperforms state-of-the-art approaches by a significant margin.

**********

Deep Rao-Blackwellised Particle Filters for Time Series Forecasting

Richard Kurle, Syama Sundar Rangapuram, Emmanuel de Bézenac, Stephan Günnemann, Jan Gasthaus

This work addresses efficient inference and learning in switching Gaussian linear dynamical systems using a Rao-Blackwellised particle filter and a corresponding Monte Carlo objective. To improve the forecasting capabilities, we extend this classical model by conditionally linear state-to-switch dynamics, while leaving the partial tractability of the conditional Gaussian linear part intact. Furthermore, we use an auxiliary variable approach with a decoder-type neural network that allows for more complex non-linear emission models and multivariate observations. We propose a Monte Carlo objective that leverages the conditional linearity by computing the corresponding conditional expectations in closed-form and a suitable proposal distribution that is factorised similarly to the optimal proposal distribution. We evaluate our approach on several popular time series forecasting datasets as well as image streams of simulated physical systems. Our results show improved forecasting performance compared to other deep state-space model approaches.

**********

Why are Adaptive Methods Good for Attention Models?

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, Suvrit Sra

While stochastic gradient descent (SGD) is still the de facto algorithm in deep learning, adaptive methods like Clipped SGD/Adam have been observed to outperform SGD across important tasks, such as attention models. The settings under which SGD performs poorly in comparison to adaptive methods are not well understood y

et. In this paper, we provide empirical and theoretical evidence that a heavy-tailed distribution of the noise in stochastic gradients is one cause of SGD's poor performance. We provide the first tight upper and lower convergence bounds for adaptive gradient methods under heavy-tailed noise. Further, we demonstrate how gradient clipping plays a key role in addressing heavy-tailed gradient noise. Subsequently, we show how clipping can be applied in practice by developing an adaptive coordinate-wise clipping algorithm (ACClip) and demonstrate its superior performance on BERT pretraining and finetuning tasks.

**********

Neural Sparse Representation for Image Restoration

Yuchen Fan, Jiahui Yu, Yiqun Mei, Yulun Zhang, Yun Fu, Ding Liu, Thomas S. Huang

Inspired by the robustness and efficiency of sparse representation in sparse coding based image restoration models, we investigate the sparsity of neurons in deep networks. Our method structurally enforces sparsity constraints upon hidden neurons. The sparsity constraints are favorable for gradient-based learning algorithms and attachable to convolution layers in various networks. Sparsity in neurons enables computation saving by only operating on non-zero components without hurting accuracy. Meanwhile, our method can magnify representation dimensionality and model capacity with negligible additional computation cost. Experiments show that sparse representation is crucial in deep neural networks for multiple image restoration tasks, including image super-resolution, image denoising, and image compression artifacts removal.

**********

Boosting First-Order Methods by Shifting Objective: New Schemes with Faster Worst-Case Rates

Kaiwen Zhou, Anthony Man-Cho So, James Cheng

We propose a new methodology to design first-order methods for unconstrained strongly convex problems. Specifically, instead of tackling the original objective directly, we construct a shifted objective function that has the same minimizer as the original objective and encodes both the smoothness and strong convexity of the original objective in an interpolation condition. We then propose an algorithmic template for tackling the shifted objective, which can exploit such a condition. Following this template, we derive several new accelerated schemes for problems that are equipped with various first-order oracles and show that the interpolation condition allows us to vastly simplify and tighten the analysis of the derived methods. In particular, all the derived methods have faster worst-case convergence rates than their existing counterparts. Experiments on machine learning tasks are conducted to evaluate the new methods.

**********

Robust Sequence Submodular Maximization

Gamal Sallam, Zizhan Zheng, Jie Wu, Bo Ji

Submodularity is an important property of set functions and has been extensively studied in the literature. It models set functions that exhibit a diminishing returns property, where the marginal value of adding an element to a set decreases as the set expands. This notion has been generalized to considering sequence functions, where the order of adding elements plays a crucial role and determines the function value; the generalized notion is called sequence (or string) submodularity. In this paper, we study a new problem of robust sequence submodular maximization with cardinality constraints. The robustness is against the removal of a subset of elements in the selected sequence (e.g., due to malfunctions or adversarial attacks). Compared to robust submodular maximization for set function, new challenges arise when sequence functions are concerned. Specifically, there are multiple definitions of submodularity for sequence functions, which exhibit subtle yet critical differences. Another challenge comes from two directions of monotonicity: forward monotonicity and backward monotonicity, both of which are important to proving performance guarantees. To address these unique challenges, we design two robust greedy algorithms: while one algorithm achieves a constant approximation ratio but is robust only against the removal of a subset of contiguous elements, the other is robust against the removal of an arbitrary subset of the selected elements but requires a stronger assumption and achieves an appr

oximation ratio that depends on the number of the removed elements. Finally, we generalize the analyses to considering sequence functions under weaker assumptions based on approximate versions of sequence submodularity and backward monotonicity.

**********

Certified Monotonic Neural Networks

Xingchao Liu, Xing Han, Na Zhang, Qiang Liu

Learning monotonic models with respect to a subset of the inputs is a desirable feature to effectively address the fairness, interpretability, and generalization issues in practice. Existing methods for learning monotonic neural networks either require specifically designed model structures to ensure monotonicity, which can be too restrictive/complicated, or enforce monotonicity by adjusting the learning process, which cannot provably guarantee the learned model is monotonic on selected features. In this work, we propose to certify the monotonicity of the general piece-wise linear neural networks by solving a mixed integer linear programming problem. This provides a new general approach for learning monotonic neural networks with arbitrary model structures. Our method allows us to train neural networks with heuristic monotonicity regularizations, and we can gradually increase the regularization magnitude until the learned network is certified monotonic. Compared to prior work, our method does not require human-designed constraints on the weight space and also yields more accurate approximation. Empirical studies on various datasets demonstrate the efficiency of our approach over the state-of-the-art methods, such as Deep Lattice Networks

**********

System Identification with Biophysical Constraints: A Circuit Model of the Inner Retina

Cornelius Schröder, David Klindt, Sarah Strauss, Katrin Franke, Matthias Bethge, Thomas Euler, Philipp Berens

Visual processing in the retina has been studied in great detail at all levels such that a comprehensive picture of the retina's cell types and the many neural circuits they form is emerging. However, the currently best performing models of retinal function are black-box CNN models which are agnostic to such biological knowledge. In particular, these models typically neglect the role of the many inhibitory circuits involving amacrine cells and the biophysical mechanisms underlying synaptic release. Here, we present a computational model of temporal processing in the inner retina, including inhibitory feedback circuits and realistic synaptic release mechanisms. Fit to the responses of bipolar cells, the model generalized well to new stimuli including natural movie sequences, performing on par with or better than a benchmark black-box model. In pharmacology experiments, the model replicated in silico the effect of blocking specific amacrine cell populations with high fidelity, indicating that it had learned key circuit functions. Also, more in depth comparisons showed that connectivity patterns learned by the model were well matched to connectivity patterns extracted from connectomics data. Thus, our model provides a biologically interpretable data-driven account of temporal processing in the inner retina, filling the gap between purely black-box and detailed biophysical modeling.

**********

Efficient Algorithms for Device Placement of DNN Graph Operators

Jakub M. Tarnawski, Amar Phanishayee, Nikhil Devanur, Divya Mahajan, Fanny Nina Paravecino

Modern machine learning workloads use large models, with complex structures, that are very expensive to execute. The devices that execute complex models are becoming increasingly heterogeneous as we see a flourishing of Domain Specific Architectures (DSAs) being offered as hardware accelerators in addition to CPUs. These trends necessitate distributing the workload across multiple devices. Recent work has shown that significant gains can be obtained with model parallelism, i.e, partitioning a neural network's computational graph onto multiple devices. In particular, this form of parallelism assumes a pipeline of devices, which is fed a stream of samples and yields high throughput for training and inference of DNNs. However, for such settings (large models and multiple heterogeneous devices

), we require automated algorithms and toolchains that can partition the ML workload across devices.
**********

Active Invariant Causal Prediction: Experiment Selection through Stability
Juan L. Gamella, Christina Heinze-Deml
A fundamental difficulty of causal learning is that causal models can generally not be fully identified based on observational data only. Interventional data, that is, data originating from different experimental environments, improves identifiability. However, the improvement depends critically on the target and nature of the interventions carried out in each experiment. Since in real applications experiments tend to be costly, there is a need to perform the right interventions such that as few as possible are required. In this work we propose a new active learning (i.e. experiment selection) framework (A-ICP) based on Invariant Causal Prediction (ICP) (Peters et al. 2016). For general structural causal models, we characterize the effect of interventions on so-called stable sets, a notion introduced by Pfister et al. 2019. We leverage these results to propose several intervention selection policies for A-ICP which quickly reveal the direct causes of a response variable in the causal graph while maintaining the error control inherent in ICP. Empirically, we analyze the performance of the proposed policies in both population and finite-regime experiments.
**********

BOSS: Bayesian Optimization over String Spaces
Henry Moss, David Leslie, Daniel Beck, Javier González, Paul Rayson
This article develops a Bayesian optimization (BO) method which acts directly over raw strings, proposing the first uses of string kernels and genetic algorithms within BO loops. Recent applications of BO over strings have been hindered by the need to map inputs into a smooth and unconstrained latent space. Learning this projection is computationally and data-intensive. Our approach instead builds a powerful Gaussian process surrogate model based on string kernels, naturally supporting variable length inputs, and performs efficient acquisition function maximization for spaces with syntactic constraints. Experiments demonstrate considerably improved optimization over existing approaches across a broad range of constraints, including the popular setting where syntax is governed by a context-free grammar.
**********

Model Interpretability through the lens of Computational Complexity
Pablo Barceló, Mikaël Monet, Jorge Pérez, Bernardo Subercaseaux
In spite of several claims stating that some models are more interpretable than others --e.g., "linear models are more interpretable than deep neural networks"-- we still lack a principled notion of interpretability that allows us to formally compare among different classes of models. We make a step towards such a theory by studying whether folklore interpretability claims have a correlate in terms of computational complexity theory. We focus on post-hoc explainability queries that, intuitively, attempt to answer why individual inputs are classified in a certain way by a given model. In a nutshell, we say that a class C1 of models is more interpretable than another class C2, if the computational complexity of answering post-hoc queries for models in C2 is higher than for C1. We prove that this notion provides a good theoretical counterpart to current beliefs on the interpretability of models; in particular, we show that under our definition and assuming standard complexity-theoretical assumptions (such as P!=NP), both linear and tree-based models are strictly more interpretable than neural networks. Our complexity analysis, however, does not provide a clear-cut difference between linear and tree-based models, as we obtain different results depending on the particular {post-hoc explanations} considered. Finally, by applying a finer complexity analysis based on parameterized complexity, we are able to prove a theoretical result suggesting that shallow neural networks are more interpretable than deeper ones.
**********

Markovian Score Climbing: Variational Inference with KL(p||q)
Christian Naesseth, Fredrik Lindsten, David Blei

Modern variational inference (VI) uses stochastic gradients to avoid intractable expectations, enabling large-scale probabilistic inference in complex models. VI posits a family of approximating distributions q and then finds the member of that family that is closest to the exact posterior p. Traditionally, VI algorithms minimize the “exclusive Kullback-Leibler (KL)” KL(q||p), often for computational convenience. Recent research, however, has also focused on the “inclusive KL” KL(p||q), which has good statistical properties that makes it more appropriate for certain inference problems. This paper develops a simple algorithm for reliably minimizing the inclusive KL using stochastic gradients with vanishing bias. This method, which we call Markovian score climbing (MSC), converges to a local optimum of the inclusive KL. It does not suffer from the systematic errors inherent in existing methods, such as Reweighted Wake-Sleep and Neural Adaptive Sequential Monte Carlo, which lead to bias in their final estimates. We illustrate convergence on a toy model and demonstrate the utility of MSC on Bayesian probit regression for classification as well as a stochastic volatility model for financial data.

**********

Improved Analysis of Clipping Algorithms for Non-convex Optimization

Bohang Zhang, Jikai Jin, Cong Fang, 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-authors prior to requesting a name change in the electronic proceedings.

**********

Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang

We develop a new approach to obtaining high probability regret bounds for online learning with bandit feedback against an adaptive adversary. While existing approaches all require carefully constructing optimistic and biased loss estimators, our approach uses standard unbiased estimators and relies on a simple increasing learning rate schedule, together with the help of logarithmically homogeneous self-concordant barriers and a strengthened Freedman's inequality.

**********

A Ranking-based, Balanced Loss Function Unifying Classification and Localisation in Object Detection

Kemal Oksuz, Baris Can Cam, Emre Akbas, Sinan Kalkan

We propose average Localisation-Recall-Precision (aLRP), a unified, bounded, balanced and ranking-based loss function for both classification and localisation tasks in object detection. aLRP extends the Localisation-Recall-Precision (LRP) performance metric (Oksuz et al., 2018) inspired from how Average Precision (AP) Loss extends precision to a ranking-based loss function for classification (Chen et al., 2020). aLRP has the following distinct advantages: (i) aLRP is the first ranking-based loss function for both classification and localisation tasks. (ii) Thanks to using ranking for both tasks, aLRP naturally enforces high-quality localisation for high-precision classification. (iii) aLRP provides provable balance between positives and negatives. (iv) Compared to on average ~6 hyperparameters in the loss functions of state-of-the-art detectors, aLRP Loss has only one hyperparameter, which we did not tune in practice. On the COCO dataset, aLRP Loss improves its ranking-based predecessor, AP Loss, up to around 5 AP points, achieves 48.9 AP without test time augmentation and outperforms all one-stage detectors. Code available at: https://github.com/kemaloksuz/aLRPLoss .

**********

StratLearner: Learning a Strategy for Misinformation Prevention in Social Networks

Guangmo Tong

Given a combinatorial optimization problem taking an input, can we learn a strategy to solve it from the examples of input-solution pairs without knowing its objective function? In this paper, we consider such a setting and study the misinf

ormation prevention problem. Given the examples of attacker-protector pairs, our goal is to learn a strategy to compute protectors against future attackers, without the need of knowing the underlying diffusion model. To this end, we design a structured prediction framework, where the main idea is to parameterize the scoring function using random features constructed through distance functions on randomly sampled subgraphs, which leads to a kernelized scoring function with weights learnable via the large margin method. Evidenced by experiments, our method can produce near-optimal protectors without using any information of the diffusion model, and it outperforms other possible graph-based and learning-based methods by an evident margin.
**********

A Unified Switching System Perspective and Convergence Analysis of Q-Learning Algorithms
Donghwan Lee, Niao He
This paper develops a novel and unified framework to analyze the convergence of a large family of Q-learning algorithms from the switching system perspective. We show that the nonlinear ODE models associated with Q-learning and many of its variants can be naturally formulated as affine switching systems. Building on their asymptotic stability, we obtain a number of interesting results: (i) we provide a simple ODE analysis for the convergence of asynchronous Q-learning under relatively weak assumptions; (ii) we establish the first convergence analysis of the averaging Q-learning algorithm; and (iii) we derive a new sufficient condition for the convergence of Q-learning with linear function approximation.
**********

Kernel Alignment Risk Estimator: Risk Prediction from Training Data
Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, Franck Gabriel
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Calibrating CNNs for Lifelong Learning
Pravendra Singh, Vinay Kumar Verma, Pratik Mazumder, Lawrence Carin, Piyush Rai
We present an approach for lifelong/continual learning of convolutional neural networks (CNN) that does not suffer from the problem of catastrophic forgetting when moving from one task to the other. We show that the activation maps generated by the CNN trained on the old task can be calibrated using very few calibration parameters, to become relevant to the new task. Based on this, we calibrate the activation maps produced by each network layer using spatial and channel-wise calibration modules and train only these calibration parameters for each new task in order to perform lifelong learning. Our calibration modules introduce significantly less computation and parameters as compared to the approaches that dynamically expand the network. Our approach is immune to catastrophic forgetting since we store the task-adaptive calibration parameters, which contain all the task-specific knowledge and is exclusive to each task. Further, our approach does not require storing data samples from the old tasks, which is done by many replay based methods. We perform extensive experiments on multiple benchmark datasets (SVHN, CIFAR, ImageNet, and MS-Celeb), all of which show substantial improvements over state-of-the-art methods (e.g., a 29% absolute increase in accuracy on CIFAR-100 with 10 classes at a time). On large-scale datasets, our approach yields 23.8% and 9.7% absolute increase in accuracy on ImageNet-100 and MS-Celeb-10K datasets, respectively, by employing very few (0.51% and 0.35% of model parameters) task-adaptive calibration parameters.
**********

Online Convex Optimization Over Erdos-Renyi Random Networks
Jinlong Lei, Peng Yi, Yiguang Hong, Jie Chen, Guodong Shi
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.

**********
Robustness of Bayesian Neural Networks to Gradient-Based Attacks
Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane', Luca Bortolussi, Guido Sanguinetti
Vulnerability to adversarial attacks is one of the principal hurdles to the adoption of deep learning in safety-critical applications. Despite significant efforts, both practical and theoretical, the problem remains open. In this paper, we analyse the geometry of adversarial attacks in the large-data, overparametrized limit for Bayesian Neural Networks (BNNs). We show that, in the limit, vulnerability to gradient-based attacks arises as a result of degeneracy in the data distribution, i.e., when the data lies on a lower-dimensional submanifold of the ambient space. As a direct consequence, we demonstrate that in the limit BNN posteriors are robust to gradient-based adversarial attacks. Experimental results on the MNIST and Fashion MNIST datasets with BNNs trained with Hamiltonian Monte Carlo and Variational Inference support this line of argument, showing that BNNs can display both high accuracy and robustness to gradient based adversarial attacks.
**********
Parametric Instance Classification for Unsupervised Visual Feature learning
Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, Han Hu
This paper presents parametric instance classification (PIC) for unsupervised visual feature learning. Unlike the state-of-the-art approaches which do instance discrimination in a dual-branch non-parametric fashion, PIC directly performs a one-branch parametric instance classification, revealing a simple framework similar to supervised classification and without the need to address the information leakage issue. We show that the simple PIC framework can be as effective as the state-of-the-art approaches, i.e. SimCLR and MoCo v2, by adapting several common component settings used in the state-of-the-art approaches. We also propose two novel techniques to further improve effectiveness and practicality of PIC: 1) a sliding-window data scheduler, instead of the previous epoch-based data scheduler, which addresses the extremely infrequent instance visiting issue in PIC and improves the effectiveness; 2) a negative sampling and weight update correction approach to reduce the training time and GPU memory consumption, which also enables application of PIC to almost unlimited training images. We hope that the PIC framework can serve as a simple baseline to facilitate future study. The code and network configurations are available at \url{https://github.com/bl0/PIC}.
**********
Sparse Weight Activation Training
Md Aamir Raihan, Tor Aamodt
Neural network training is computationally and memory intensive. Sparse training can reduce the burden on emerging hardware platforms designed to accelerate sparse computations, but it can also affect network convergence. In this work, we propose a novel CNN training algorithm called Sparse Weight Activation Training (SWAT). SWAT is more computation and memory-efficient than
conventional training. SWAT modifies back-propagation based on the empirical insight that convergence during training tends to be robust to the elimination of (i) small magnitude weights during the forward pass and (ii) both small magnitude weights and activations during the backward pass. We evaluate SWAT on recent CNN architectures such as ResNet, VGG, DenseNet and WideResNet using CIFAR-10, CIFAR-100 and ImageNet datasets. For ResNet-50 on ImageNet SWAT reduces total floating-point operations (FLOPs) during training by 80% resulting in a 3.3x training speedup when run on a simulated sparse learning accelerator representative of emerging platforms while incurring only 1.63% reduction in validation accuracy. Moreover, SWAT reduces memory footprint during the backward pass by 23% to 50% for activations and 50% to 90% for weights. Code is available at https://github.com/AamirRaihan/SWAT.
**********
Collapsing Bandits and Their Application to Public Health Intervention
Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, Milind Tambe
We propose and study Collapsing Bandits,  a new restless multi-armed bandit (RMA

B) setting in which each arm follows a binary-state Markovian process with a special structure: when an arm is played, the state is fully observed, thus“collapsing” any uncertainty, but when an arm is passive, no observation is made, thus allowing uncertainty to evolve. The goal is to keep as many arms in the “good” state as possible by planning a limited budget of actions per round. Such CollapsingBandits are natural models for many healthcare domains in which health workers must simultaneously monitor patients and deliver interventions in a way that maximizes the health of their patient cohort. Our main contributions are as follows: (i) Building on the Whittle index technique for RMABs, we derive conditions under which the Collapsing Bandits problem is indexable. Our derivation hinges on novel conditions that characterize when the optimal policies may take the form of either“forward” or “reverse” threshold policies. (ii) We exploit the optimality of threshold policies to build fast algorithms for computing the Whittle index, including a closed-form. (iii) We evaluate our algorithm on several data distributions including data from a real-world healthcare task in which a worker must monitor and deliver interventions to maximize their patients’ adherence to tuberculosis medication. Our algorithm achieves a 3-order-of-magnitude speedup compared to state-of-the-art RMAB techniques, while achieving similar performance. The code is available at:https://github.com/AdityaMate/collapsing_bandits
**********

Neural Sparse Voxel Fields

Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, Christian Theobalt

Photo-realistic free-viewpoint rendering of real-world scenes using classical computer graphics techniques is challenging, because it requires the difficult step of capturing detailed appearance and geometry models. Recent studies have demonstrated promising results by learning scene representations that implicitly encodes both geometry and appearance without 3D supervision. However, existing approaches in practice often show blurry renderings caused by the limited network capacity or the difficulty in finding accurate intersections of camera rays with the scene geometry. Synthesizing high-resolution imagery from these representations often requires time-consuming optical ray marching. In this work, we introduce Neural Sparse Voxel Fields (NSVF), a new neural scene representation for fast and high-quality free-viewpoint rendering. The NSVF defines a series of voxel-bounded implicit fields organized in a sparse voxel octree to model local properties in each cell. We progressively learn the underlying voxel structures with a differentiable ray-marching operation from only a set of posed RGB images. With the sparse voxel octree structure, rendering novel views at inference time can be accelerated by skipping the voxels without relevant scene content. Our method is over 10 times faster than the state-of-the-art while achieving higher quality results. Furthermore, by utilizing an explicit sparse voxel representation, our method can be easily applied to scene editing and scene composition. we also demonstrate various kinds of challenging tasks, including multi-object learning, free-viewpoint rendering of a moving human, and large-scale scene rendering.
**********

A Flexible Framework for Designing Trainable Priors with Adaptive Smoothing and Game Encoding

Bruno Lecouat, Jean Ponce, Julien Mairal

We introduce a general framework for designing and training neural network layers whose forward passes can be interpreted as solving non-smooth convex optimization problems, and whose architectures are derived from an optimization algorithm. We focus on convex games, solved by local agents represented by the nodes of a graph and interacting through regularization functions. This approach is appealing for solving imaging problems, as it allows the use of classical image priors within deep models that are trainable end to end. The priors used in this presentation include variants of total variation, Laplacian regularization, bilateral filtering, sparse coding on learned dictionaries, and non-local self similarities. Our models are fully interpretable as well as parameter and data efficient. Our experiments demonstrate their effectiveness on a large diversity of tasks ranging from image denoising and compressed sensing for fMRI to dense stereo matching.

**********
The Discrete Gaussian for Differential Privacy
Clément L. Canonne, Gautam Kamath, Thomas Steinke
A key tool for building differentially private systems is adding Gaussian noise to the output of a function evaluated on a sensitive dataset. Unfortunately, using a continuous distribution presents several practical challenges.
First and foremost, finite computers cannot exactly represent samples from continuous distributions, and previous work has demonstrated that seemingly innocuous numerical errors can entirely destroy privacy. Moreover, when the underlying data is itself discrete (e.g., population counts), adding continuous noise makes the result less interpretable.
**********
Robust Sub-Gaussian Principal Component Analysis and Width-Independent Schatten Packing
Arun Jambulapati, Jerry Li, 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.
**********
Adaptive Importance Sampling for Finite-Sum Optimization and Sampling with Decreasing Step-Sizes
Ayoub El Hanchi, David Stephens
Reducing the variance of the gradient estimator is known to improve the convergence rate of stochastic gradient-based optimization and sampling algorithms. One way of achieving variance reduction is to design importance sampling strategies. Recently, the problem of designing such schemes was formulated as an online learning problem with bandit feedback, and algorithms with sub-linear static regret were designed. In this work, we build on this framework and propose a simple and efficient algorithm for adaptive importance sampling for finite-sum optimization and sampling with decreasing step-sizes. Under standard technical conditions, we show that our proposed algorithm achieves O(T^{2/3}) and O(T^{5/6}) dynamic regret for SGD and SGLD respectively when run with O(1/t) step sizes. We achieve this dynamic regret bound by leveraging our knowledge of the dynamics defined by the algorithm, and combining ideas from online learning and variance-reduced stochastic optimization. We validate empirically the performance of our algorithm and identify settings in which it leads to significant improvements.
**********
Learning efficient task-dependent representations with synaptic plasticity
Colin Bredenberg, Eero Simoncelli, Cristina Savin
Neural populations encode the sensory world imperfectly: their capacity is limited by the number of neurons, availability of metabolic and other biophysical resources, and intrinsic noise. The brain is presumably shaped by these limitations, improving efficiency by discarding some aspects of incoming sensory streams, while preferentially preserving commonly occurring, behaviorally-relevant information. Here we construct a stochastic recurrent neural circuit model that can learn efficient, task-specific sensory codes using a novel form of reward-modulated Hebbian synaptic plasticity. We illustrate the flexibility of the model by training an initially unstructured neural network to solve two different tasks: stimulus estimation, and stimulus discrimination. The network achieves high performance in both tasks by appropriately allocating resources and using its recurrent circuitry to best compensate for different levels of noise. We also show how the interaction between stimulus priors and task structure dictates the emergent network representations.
**********
A Contour Stochastic  Gradient Langevin Dynamics Algorithm for Simulations of Multi-modal Distributions
Wei Deng, Guang Lin, Faming Liang
We propose an adaptively weighted stochastic gradient Langevin dynamics algorithm (SGLD), so-called contour stochastic gradient Langevin dynamics (CSGLD), for B

ayesian learning in big data statistics. The proposed algorithm is essentially a scalable dynamic importance sampler, which automatically flattens the target distribution such that the simulation for a multi-modal distribution can be greatly facilitated. Theoretically, we prove a stability condition and establish the asymptotic convergence of the self-adapting parameter to a unique fixed-point, regardless of the non-convexity of the original energy function; we also present an error analysis for the weighted averaging estimators. Empirically, the CSGLD algorithm is tested on multiple benchmark datasets including CIFAR10 and CIFAR100. The numerical results indicate its superiority over the existing state-of-the-art algorithms in training deep neural networks.
**********

Error Bounds of Imitating Policies and Environments
Tian Xu, Ziniu Li, Yang Yu
Imitation learning trains a policy by mimicking expert demonstrations. Various imitation methods were proposed and empirically evaluated, meanwhile, their theoretical understanding needs further studies. In this paper, we firstly analyze the value gap between the expert policy and imitated policies by two imitation methods, behavioral cloning and generative adversarial imitation. The results support that generative adversarial imitation can reduce the compounding errors compared to behavioral cloning, and thus has a better sample complexity. Noticed that by considering the environment transition model as a dual agent, imitation learning can also be used to learn the environment model. Therefore, based on the bounds of imitating policies, we further analyze the performance of imitating environments. The results show that environment models can be more effectively imitated by generative adversarial imitation than behavioral cloning, suggesting a novel application of adversarial imitation for model-based reinforcement learning. We hope these results could inspire future advances in imitation learning and model-based reinforcement learning.
**********

Disentangling Human Error from Ground Truth in Segmentation of Medical Images
Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, Daniel Alexander
Recent years have seen increasing use of supervised learning methods for segmentation tasks. However, the predictive performance of these algorithms depends on the quality of labels. This problem is particularly pertinent in the medical image domain, where both the annotation cost and inter-observer variability are high. In a typical label acquisition process, different human experts provide their estimates of the ``true'' segmentation labels under the influence of their own biases and competence levels. Treating these noisy labels blindly as the ground truth limits the performance that automatic segmentation algorithms can achieve. In this work, we present a method for jointly learning, from purely noisy observations alone, the reliability of individual annotators and the true segmentation label distributions, using two coupled CNNs. The separation of the two is achieved by encouraging the estimated annotators to be maximally unreliable while achieving high fidelity with the noisy training data. We first define a toy segmentation dataset based on MNIST and study the properties of the proposed algorithm. We then demonstrate the utility of the method on three public medical imaging segmentation datasets with simulated (when necessary) and real diverse annotations: 1) MSLSC (multiple-sclerosis lesions); 2) BraTS (brain tumours); 3) LIDC-IDRI (lung abnormalities). In all cases, our method outperforms competing methods and relevant baselines particularly in cases where the number of annotations is small and the amount of disagreement is large. The experiments also show strong ability to capture the complex spatial characteristics of annotators' mistakes. Our code is available at \url{https://github.com/moucheng2017/LearnNoisyLabelsMedicalImages}.
**********

Consequences of Misaligned AI
Simon Zhuang, Dylan Hadfield-Menell
AI systems often rely on two key components: a specified goal or reward function and an optimization algorithm to compute the optimal behavior for that

goal. This approach is intended to provide value for a principal: the user on whose behalf the agent acts. The objectives given to these agents often refer to a partial specification of the principal's goals. We consider the cost of this incompleteness by analyzing a model of a principal and an agent in a resource constrained world where the L features of the state correspond to different sources of utility for the principal. We assume that the reward function given to the agent only has support on J < L features. The contributions of our paper are as follows: 1) we propose a novel model of an incomplete principal—agent problem from artificial intelligence; 2) we provide necessary and sufficient conditions under which indefinitely optimizing for any incomplete proxy objective leads to arbitrarily low overall utility; and 3) we show how modifying the setup to allow reward functions that reference the full state or allowing the principal to update the proxy objective over time can lead to higher utility solutions. The results in this paper argue that we should view the design of reward functions as an interactive and dynamic process and identifies a theoretical scenario where some degree of interactivity is desirable.
**********

Promoting Coordination through Policy Regularization in Multi-Agent Deep Reinforcement Learning

Julien Roy, Paul Barde, Félix Harvey, Derek Nowrouzezahrai, Chris Pal

In multi-agent reinforcement learning, discovering successful collective behaviors is challenging as it requires exploring a joint action space that grows exponentially with the number of agents. While the tractability of independent agent-wise exploration is appealing, this approach fails on tasks that require elaborate group strategies. We argue that coordinating the agents' policies can guide their exploration and we investigate techniques to promote such an inductive bias. We propose two policy regularization methods: TeamReg, which is based on inter-agent action predictability and CoachReg that relies on synchronized behavior selection. We evaluate each approach on four challenging continuous control tasks with sparse rewards that require varying levels of coordination as well as on the discrete action Google Research Football environment. Our experiments show improved performance across many cooperative multi-agent problems. Finally, we analyze the effects of our proposed methods on the policies that our agents learn and show that our methods successfully enforce the qualities that we propose as proxies for coordinated behaviors.
**********

Emergent Reciprocity and Team Formation from Randomized Uncertain Social Preferences

Bowen Baker

Multi-agent reinforcement learning (MARL) has shown recent success in increasingly complex fixed-team zero-sum environments. However, the real world is not zero-sum nor does it have fixed teams; humans face numerous social dilemmas and must learn when to cooperate and when to compete. To successfully deploy agents into the human world, it may be important that they be able to understand and help in our conflicts. Unfortunately, selfish MARL agents typically fail when faced with social dilemmas. In this work, we show evidence of emergent direct reciprocity, indirect reciprocity and reputation, and team formation when training agents with randomized uncertain social preferences (RUSP), a novel environment augmentation that expands the distribution of environments agents play in. RUSP is generic and scalable; it can be applied to any multi-agent environment without changing the original underlying game dynamics or objectives. In particular, we show that with RUSP these behaviors can emerge and lead to higher social welfare equilibria in both classic abstract social dilemmas like Iterated Prisoner's Dilemma as well in more complex intertemporal environments.
**********

Hitting the High Notes: Subset Selection for Maximizing Expected Order Statistics

Aranyak Mehta, Uri Nadav, Alexandros Psomas, Aviad Rubinstein

Requests for name changes in the electronic proceedings will be accepted with 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 Scale-Invariant Graph-related Problem Solving by Iterative Homogeneous G
NNs

Hao Tang, Zhiao Huang, Jiayuan Gu, Bao-Liang Lu, Hao Su

Current graph neural networks (GNNs) lack generalizability with respect to scale
s (graph sizes, graph diameters, edge weights, etc..) when solving many graph an
alysis problems. Taking the perspective of synthesizing graph theory programs, w
e propose several extensions to address the issue. First, inspired by the depend
ency of iteration number of common  graph theory algorithms on graph size, we le
arn to terminate the message passing process in GNNs adaptively according to the
 computation progress. Second, inspired by the fact that many graph theory algor
ithms are homogeneous with respect to graph weights, we introduce homogeneous tr
ansformation layers that are universal homogeneous function approximators, to co
nvert ordinary GNNs to be homogeneous. Experimentally, we show that our GNN can
be trained from small-scale graphs but generalize well to large-scale graphs for
 a number of basic graph theory problems. It also shows generalizability for app
lications of multi-body physical simulation and image-based navigation problems.
**********

Regret Bounds without Lipschitz Continuity: Online Learning with Relative-Lipsch
itz Losses

Yihan Zhou, Victor Sanches Portella, Mark Schmidt, Nicholas Harvey

In online convex optimization (OCO), Lipschitz continuity of the functions is co
mmonly assumed in order to obtain sublinear regret. Moreover, many algorithms ha
ve only logarithmic regret when these functions are also strongly convex. Recent
ly, researchers from convex optimization proposed the notions of relative Lipsch
itz continuity'' andrelative strong convexity''. Both of the notions are general
izations of their classical counterparts. It has been shown that subgradient met
hods in the relative setting have performance analogous to their performance in
the classical setting.
**********

The Lottery Ticket Hypothesis for Pre-trained BERT Networks

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang W
ang, Michael Carbin

In natural language processing (NLP), enormous pre-trained models like BERT have
 become the standard starting point for training on a range of downstream tasks,
 and similar trends are emerging in other areas of deep learning. In parallel, w
ork on the lottery ticket hypothesis has shown that models for NLP and computer
vision contain smaller matching subnetworks capable of training in isolation to
full accuracy and transferring to other tasks. In this work, we combine these ob
servations to assess whether such trainable, transferrable subnetworks exist in
pre-trained BERT models. For a range of downstream tasks, we indeed find matchin
g subnetworks at 40% to 90% sparsity. We find these subnetworks at (pre-trained)
 initialization, a deviation from prior NLP research where they emerge only afte
r some amount of training. Subnetworks found on the masked language modeling tas
k (the same task used to pre-train the model) transfer universally; those found
on other tasks transfer in a limited fashion if at all. As large-scale pre-train
ing becomes an increasingly central paradigm in deep learning, our results demon
strate that the main lottery ticket observations remain relevant in this context
. Codes available at https://github.com/VITA-Group/BERT-Tickets.
**********

Label-Aware Neural Tangent Kernel: Toward Better Generalization and Local Elasti
city

Shuxiao Chen, Hangfeng He, Weijie Su

As a popular approach to modeling the dynamics of training overparametrized neur
al networks (NNs), the neural tangent kernels (NTK) are known to fall behind rea
l-world NNs in generalization ability. This performance gap is in part due to th
e \textit{label agnostic} nature of the NTK, which renders the resulting kernel
not as \textit{locally elastic} as NNs~\citep{he2019local}. In this paper, we in

troduce a novel approach from the perspective of \emph{label-awareness} to reduc
e this gap for the NTK. Specifically, we propose two label-aware kernels that ar
e each a superimposition of a label-agnostic part and a hierarchy of label-aware
 parts with increasing complexity of label dependence, using the Hoeffding decom
position. Through both theoretical and empirical evidence, we show that the mode
ls trained with the proposed kernels better simulate NNs in terms of generalizat
ion ability and local elasticity.
**********

Beyond Perturbations: Learning Guarantees with Arbitrary Adversarial Test Exampl
es

Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, Omar Montasser

We present a transductive learning algorithm that takes as input training exampl
es
from a distribution P and arbitrary (unlabeled) test examples, possibly chosen b
y
an adversary. This is unlike prior work that assumes that test examples are smal
l
perturbations of P. Our algorithm outputs a selective classifier, which abstains
 from predicting on some examples. By considering selective transductive learnin
g, we give the first nontrivial guarantees for learning classes of bounded VC di
mension with arbitrary train and test distributions—no prior guarantees were kno
wn even for simple classes of functions such as intervals on the line. In partic
ular, for any function in a class C of bounded VC dimension, we guarantee a low
test error rate and a low rejection rate with respect to P. Our algorithm is eff
icient given an Empirical Risk Minimizer (ERM) for C. Our guarantees hold even f
or test examples chosen by an unbounded white-box adversary. We also give guaran
tees for generalization, agnostic, and unsupervised settings.
**********

AdvFlow: Inconspicuous Black-box Adversarial Attacks using Normalizing Flows

Hadi Mohaghegh Dolatabadi, Sarah Erfani, Christopher Leckie

Deep learning classifiers are susceptible to well-crafted, imperceptible variati
ons of their inputs, known as adversarial attacks. In this regard, the study of
powerful attack models sheds light on the sources of vulnerability in these clas
sifiers, hopefully leading to more robust ones. In this paper, we introduce AdvF
low: a novel black-box adversarial attack method on image classifiers that explo
its the power of normalizing flows to model the density of adversarial examples
around a given target image. We see that the proposed method generates adversari
es that closely follow the clean data distribution, a property which makes their
 detection less likely. Also, our experimental results show competitive performa
nce of the proposed approach with some of the existing attack methods on defende
d classifiers.
**********

Few-shot Image Generation with Elastic Weight Consolidation

Yijun Li, Richard Zhang, Jingwan (Cynthia) Lu, Eli Shechtman

Few-shot image generation seeks to generate more data of a given domain, with on
ly few available training examples. As it is unreasonable to expect to fully inf
er the distribution from just a few observations (e.g., emojis), we seek to leve
rage a large, related source domain as pretraining (e.g., human faces). Thus, we
 wish to preserve the diversity of the source domain, while adapting to the appe
arance of the target. We adapt a pretrained model, without introducing any addit
ional parameters, to the few examples of the target domain. Crucially, we regula
rize the changes of the weights during this adaptation, in order to best preserv
e the information of the source dataset, while fitting the target. We demonstrat
e the effectiveness of our algorithm by generating high-quality results of diffe
rent target domains, including those with extremely few examples (e.g., 10). We
also analyze the performance of our method with respect to some important factor
s, such as the number of examples and the similarity between the source and targ
et domain.
**********

On the Expressiveness of Approximate Inference in Bayesian Neural Networks

Andrew Foong, David Burt, Yingzhen Li, Richard Turner

While Bayesian neural networks (BNNs) hold the promise of being flexible, well-calibrated statistical models, inference often requires approximations whose consequences are poorly understood. We study the quality of common variational methods in approximating the Bayesian predictive distribution. For single-hidden layer ReLU BNNs, we prove a fundamental limitation in function-space of two of the most commonly used distributions defined in weight-space: mean-field Gaussian and Monte Carlo dropout. We find there are simple cases where neither method can have substantially increased uncertainty in between well-separated regions of low uncertainty. We provide strong empirical evidence that exact inference does not have this pathology, hence it is due to the approximation and not the model. In contrast, for deep networks, we prove a universality result showing that there exist approximate posteriors in the above classes which provide flexible uncertainty estimates. However, we find empirically that pathologies of a similar form as in the single-hidden layer case can persist when performing variational inference in deeper networks. Our results motivate careful consideration of the implications of approximate inference methods in BNNs.

**********

Non-Crossing Quantile Regression for Distributional Reinforcement Learning

Fan Zhou, Jianing Wang, Xingdong Feng

Distributional reinforcement learning (DRL) estimates the distribution over future returns instead of the mean to more efficiently capture the intrinsic uncertainty of MDPs. However, batch-based DRL algorithms cannot guarantee the non-decreasing property of learned quantile curves especially at the early training stage, leading to abnormal distribution estimates and reduced model interpretability. To address these issues, we introduce a general DRL framework by using non-crossing quantile regression to ensure the monotonicity constraint within each sampled batch, which can be incorporated with any well-known DRL algorithm. We demonstrate the validity of our method from both the theory and model implementation perspectives. Experiments on Atari 2600 Games show that some state-of-art DRL algorithms with the non-crossing modification can significantly outperform their baselines in terms of faster convergence speeds and better testing performance. In particular, our method can effectively recover the distribution information and thus dramatically increase the exploration efficiency when the reward space is extremely sparse.

**********

Dark Experience for General Continual Learning: a Strong, Simple Baseline

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, SIMONE CALDERARA

Continual Learning has inspired a plethora of approaches and evaluation settings; however, the majority of them overlooks the properties of a practical scenario, where the data stream cannot be shaped as a sequence of tasks and offline training is not viable. We work towards General Continual Learning (GCL), where task boundaries blur and the domain and class distributions shift either gradually or suddenly. We address it through mixing rehearsal with knowledge distillation and regularization; our simple baseline, Dark Experience Replay, matches the network's logits sampled throughout the optimization trajectory, thus promoting consistency with its past. By conducting an extensive analysis on both standard benchmarks and a novel GCL evaluation setting (MNIST-360), we show that such a seemingly simple baseline outperforms consolidated approaches and leverages limited resources. We further explore the generalization capabilities of our objective, showing its regularization being beneficial beyond mere performance.

**********

Learning to Utilize Shaping Rewards: A New Approach of Reward Shaping

Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, Changjie Fan

Reward shaping is an effective technique for incorporating domain knowledge into reinforcement learning (RL). Existing approaches such as potential-based reward shaping normally make full use of a given shaping reward function. However, since the transformation of human knowledge into numeric reward values is often imperfect due to reasons such as human cognitive bias, completely utilizing the sha

ping reward function may fail to improve the performance of RL algorithms. In th is paper, we consider the problem of adaptively utilizing a given shaping reward function. We formulate the utilization of shaping rewards as a bi-level optimiz ation problem, where the lower level is to optimize policy using the shaping rew ards and the upper level is to optimize a parameterized shaping weight function for true reward maximization. We formally derive the gradient of the expected tr ue reward with respect to the shaping weight function parameters and accordingly propose three learning algorithms based on different assumptions. Experiments i n sparse-reward cartpole and MuJoCo environments show that our algorithms can fu lly exploit beneficial shaping rewards, and meanwhile ignore unbeneficial shapin g rewards or even transform them into beneficial ones.

**********

Neural encoding with visual attention

Meenakshi Khosla, Gia Ngo, Keith Jamison, Amy Kuceyeski, Mert Sabuncu

Visual perception is critically influenced by the focus of attention. Due to lim ited resources, it is well known that neural representations are biased in favor of attended locations. Using concurrent eye-tracking and functional Magnetic Re sonance Imaging (fMRI) recordings from a large cohort of human subjects watching movies, we first demonstrate that leveraging gaze information, in the form of a ttentional masking, can significantly improve brain response prediction accuracy in a neural encoding model. Next, we propose a novel approach to neural encodin g by including a trainable soft-attention module. Using our new approach, we dem onstrate that it is possible to learn visual attention policies by end-to-end le arning merely on fMRI response data, and without relying on any eye-tracking. Interestingly, we find that attention locations estimated by the model on indepe ndent data agree well with the corresponding eye fixation patterns, despite no e xplicit supervision to do so. Together, these findings suggest that attention mo dules can be instrumental in neural encoding models of visual stimuli.

**********

On the linearity of large non-linear models: when and why the tangent kernel is constant

Chaoyue Liu, Libin Zhu, Misha Belkin

The goal of this work is to shed light on the remarkable phenomenon of "transiti on to linearity" of certain neural networks as their width approaches infinity. We show that the "transition to linearity'' of the model and, equivalently, cons tancy of the (neural) tangent kernel (NTK) result from the scaling properties of the norm of the Hessian matrix of the network as a function of the network widt h.

We present a general framework for understanding the constancy of the tangent ke rnel via Hessian scaling applicable to the standard classes of neural networks. Our analysis provides a new perspective on the phenomenon of constant tangent ke rnel, which is different from the widely accepted "lazy training''.

Furthermore, we show that the "transition to linearity" is not a general propert y of wide neural networks and does not hold when the last layer of the network i s non-linear.

It is also not necessary for successful optimization by gradient descent.

**********

PLLay: Efficient Topological Layer based on Persistent Landscapes

Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Kim, Frederic Chazal, Larry Wasserman

We propose PLLay, a novel topological layer for general deep learning models bas ed on persistence landscapes, in which we can efficiently exploit the underlying topological features of the input data structure. In this work, we show differe ntiability with respect to layer inputs, for a general persistent homology with arbitrary filtration. Thus, our proposed layer can be placed anywhere in the net work and feed critical information on the topological features of input data int o subsequent layers to improve the learnability of the networks toward a given t ask. A task-optimal structure of PLLay is learned during training via backpropag ation, without requiring any input featurization or data preprocessing. We provi de a novel adaptation for the DTM function-based filtration, and show that the p roposed layer is robust against noise and outliers through a stability analysis.

We demonstrate the effectiveness of our approach by classification experiments on various datasets.
**********
Decentralized Langevin Dynamics for Bayesian Learning
Anjaly Parayil, He Bai, Jemin George, Prudhvi Gurram
Motivated by decentralized approaches to machine learning, we propose a collaborative Bayesian learning algorithm taking the form of decentralized Langevin dynamics in a non-convex setting. Our analysis show that the initial KL-divergence between the Markov Chain and the target posterior distribution is exponentially decreasing while the error contributions to the overall KL-divergence from the additive noise is decreasing in polynomial time. We further show that the polynomial-term experiences speed-up with number of agents and provide sufficient conditions on the time-varying step-sizes to guarantee convergence to the desired distribution. The performance of the proposed algorithm is evaluated on a wide variety of machine learning tasks. The empirical results show that the performance of individual agents with locally available data is on par with the centralized setting with considerable improvement in the convergence rate.
**********
Shared Space Transfer Learning for analyzing multi-site fMRI data
Tony Muhammad Yousefnezhad, Alessandro Selvitella, Daoqiang Zhang, Andrew Greenshaw, Russell Greiner
Multi-voxel pattern analysis (MVPA) learns predictive models from task-based functional magnetic resonance imaging (fMRI) data, for distinguishing when subjects are performing different cognitive tasks — e.g., watching movies or making decisions. MVPA works best with a well-designed feature set and an adequate sample size. However, most fMRI datasets are noisy, high-dimensional, expensive to collect, and with small sample sizes. Further, training a robust, generalized predictive model that can analyze homogeneous cognitive tasks provided by multi-site fMRI datasets has additional challenges. This paper proposes the Shared Space Transfer Learning (SSTL) as a novel transfer learning (TL) approach that can functionally align homogeneous multi-site fMRI datasets, and so improve the prediction performance in every site. SSTL first extracts a set of common features for all subjects in each site. It then uses TL to map these site-specific features to a site-independent shared space in order to improve the performance of the MVPA. SSTL uses a scalable optimization procedure that works effectively for high-dimensional fMRI datasets. The optimization procedure extracts the common features for each site by using a single-iteration algorithm and maps these site-specific common features to the site-independent shared space. We evaluate the effectiveness of the proposed method for transferring between various cognitive tasks. Our comprehensive experiments validate that SSTL achieves superior performance to other state-of-the-art analysis techniques.
**********
The Diversified Ensemble Neural Network
Shaofeng Zhang, Meng Liu, Junchi Yan
Ensemble is a general way of improving the accuracy and stability of learning models, especially for the generalization ability on small datasets. Compared with tree-based methods, relatively less works have been devoted to an in-depth study on effective ensemble design for neural networks. In this paper, we propose a principled ensemble technique by constructing the so-called diversified ensemble layer to combine multiple networks as individual modules. We theoretically show that each individual model in our ensemble layer corresponds to weights in the ensemble layer optimized in different directions. Meanwhile, the devised ensemble layer can be readily integrated into popular neural architectures, including CNNs, RNNs, and GCNs. Extensive experiments are conducted on public tabular datasets, images, and texts. By adopting weight sharing approach, the results show our method can notably improve the accuracy and stability of the original neural networks with ignorable extra time and space overhead.
**********
Inductive Quantum Embedding
Santosh Kumar Srivastava, Dinesh Khandelwal, Dhiraj Madan, Dinesh Garg, Hima Kar

anam, L Venkata Subramaniam
Quantum logic inspired embedding (aka Quantum Embedding (QE)) of a Knowledge-Base (KB) was proposed recently by Garg:2019. It is claimed that the QE preserves the logical structure of the input KB given in the form of unary and binary predicates hierarchy. Such structure preservation allows one to perform Boolean logic style deductive reasoning directly over these embedding vectors. The original QE idea, however, is limited to the transductive (not inductive) setting. Moreover, the original QE scheme runs quite slow on real applications involving millions of entities. This paper alleviates both of these key limitations. We start by reformulating the original QE problem to allow for the induction. On the way, we also underscore some interesting analytic and geometric properties of the solution and leverage them to design a faster training scheme. As an application, we show that one can achieve state-of-the-art performance on the well-known NLP task of fine-grained entity type classification by using the inductive QE approach. Our training runs 9-times faster than the original QE scheme on this task.
**********
Variational Bayesian Unlearning
Quoc Phong Nguyen, Bryan Kian Hsiang Low, Patrick Jaillet
This paper studies the problem of approximately unlearning a Bayesian model from a small subset of the training data to be erased. We frame this problem as one of minimizing the Kullback-Leibler divergence between the approximate posterior belief of model parameters after directly unlearning from erased data vs. the exact posterior belief from retraining with remaining data. Using the variational inference (VI) framework, we show that it is equivalent to minimizing an evidence upper bound which trades off between fully unlearning from erased data vs. not entirely forgetting the posterior belief given the full data (i.e., including the remaining data); the latter prevents catastrophic unlearning that can render the model useless. In model training with VI, only an approximate (instead of exact) posterior belief given the full data can be obtained, which makes unlearning even more challenging. We propose two novel tricks to tackle this challenge. We empirically demonstrate our unlearning methods on Bayesian models such as sparse Gaussian process and logistic regression using synthetic and real-world datasets.
**********
Batched Coarse Ranking in Multi-Armed Bandits
Nikolai Karpov, Qin 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.
**********
Understanding and Improving Fast Adversarial Training
Maksym Andriushchenko, Nicolas Flammarion
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Coded Sequential Matrix Multiplication For Straggler Mitigation
Nikhil Krishnan Muralee Krishnan, Seyederfan Hosseini, Ashish Khisti
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Attack of the Tails: Yes, You Really Can Backdoor Federated Learning
Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, Dimitris Papailiopoulos
Due to its decentralized nature, Federated Learning (FL) lends itself to adversarial attacks in the form of backdoors during training. The goal of a backdoor is

to corrupt the performance of the trained model on specific sub-tasks (e.g., by classifying green cars as frogs). A range of FL backdoor attacks have been introduced in the literature, but also methods to defend against them, and it is currently an open question whether FL systems can be tailored to be robust against backdoors. In this work, we provide evidence to the contrary. We first establish that, in the general case, robustness to backdoors implies model robustness to adversarial examples, a major open problem in itself. Furthermore, detecting the presence of a backdoor in a FL model is unlikely assuming first-order oracles or polynomial time. We couple our theoretical results with a new family of backdoor attacks, which we refer to as edge-case backdoors. An edge-case backdoor forces a model to misclassify on seemingly easy inputs that are however unlikely to be part of the training, or test data, i.e., they live on the tail of the input distribution. We explain how these edge-case backdoors can lead to unsavory failures and may have serious repercussions on fairness. We further exhibit that, with careful tuning at the side of the adversary, one can insert them across a range of machine learning tasks (e.g., image classification, OCR, text prediction, sentiment analysis), and bypass state-of-the-art defense mechanisms.
**********

Certifiably Adversarially Robust Detection of Out-of-Distribution Data
Julian Bitterwolf, Alexander Meinke, 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.
**********

Domain Generalization via Entropy Regularization
Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, Dacheng Tao
Domain generalization aims to learn from multiple source domains a predictive model that can generalize to unseen target domains. One essential problem in domain generalization is to learn discriminative domain-invariant features. To arrive at this, some methods introduce a domain discriminator through adversarial learning to match the feature distributions in multiple source domains. However, adversarial training can only guarantee that the learned features have invariant marginal distributions, while the invariance of conditional distributions is more important for prediction in new domains. To ensure the conditional invariance of learned features, we propose an entropy regularization term that measures the dependency between the learned features and the class labels. Combined with the typical task-related loss, e.g., cross-entropy loss for classification, and adversarial loss for domain discrimination, our overall objective is guaranteed to learn conditional-invariant features across all source domains and thus can learn classifiers with better generalization capabilities.
We demonstrate the effectiveness of our method through comparison with state-of-the-art methods on both simulated and real-world datasets. Code is available at: https://github.com/sshan-zhao/DGviaER.
**********

Bayesian Meta-Learning for the Few-Shot Setting via Deep Kernels
Massimiliano Patacchiola, Jack Turner, Elliot J. Crowley, Michael O'Boyle, Amos J. Storkey
Recently, different machine learning methods have been introduced to tackle the challenging few-shot learning scenario that is, learning from a small labeled dataset related to a specific task. Common approaches have taken the form of meta-learning: learning to learn on the new problem given the old. Following the recognition that meta-learning is implementing learning in a multi-level model, we present a Bayesian treatment for the meta-learning inner loop through the use of deep kernels. As a result we can learn a kernel that transfers to new tasks; we call this Deep Kernel Transfer (DKT). This approach has many advantages: is straightforward to implement as a single optimizer, provides uncertainty quantification, and does not require estimation of task-specific parameters. We empirically demonstrate that DKT outperforms several state-of-the-art algorithms in few-shot classification, and is the state of the art for cross-domain adaptation and re

gression. We conclude that complex meta-learning routines can be replaced by a s impler Bayesian model without loss of accuracy.
**********

Skeleton-bridged Point Completion: From Global Inference to Local Adjustment
Yinyu Nie, Yiqun Lin, Xiaoguang Han, Shihui Guo, Jian Chang, Shuguang Cui, Jian. J Zhang
Point completion refers to complete the missing geometries of objects from parti al point clouds. Existing works usually estimate the missing shape by decoding a latent feature encoded from the input points. However, real-world objects are u sually with diverse topologies and surface details, which a latent feature may f ail to represent to recover a clean and complete surface. To this end, we propos e a skeleton-bridged point completion network (SK-PCN) for shape completion. Giv en a partial scan, our method first predicts its 3D skeleton to obtain the globa l structure, and completes the surface by learning displacements from skeletal p oints. We decouple the shape completion into structure estimation and surface re construction, which eases the learning difficulty and benefits our method to obt ain on-surface details. Besides, considering the missing features during encodin g input points, SK-PCN adopts a local adjustment strategy that merges the input point cloud to our predictions for surface refinement. Comparing with previous m ethods, our skeleton-bridged manner better supports point normal estimation to o btain the full surface mesh beyond point clouds. The qualitative and quantitativ e experiments on both point cloud and mesh completion show that our approach out performs the existing methods on various object categories.
**********

Compressing Images by Encoding Their Latent Representations with Relative Entrop y Coding
Gergely Flamich, Marton Havasi, José Miguel Hernández-Lobato
Variational Autoencoders (VAEs) have seen widespread use in learned image compre ssion. They are used to learn expressive latent representations on which downstr eam compression methods can operate with high efficiency. Recently proposed 'bit s-back' methods can indirectly encode the latent representation of images with c odelength close to the relative entropy between the latent posterior and the pri or. However, due to the underlying algorithm, these methods can only be used for lossless compression, and they only achieve their nominal efficiency when compr essing multiple images simultaneously; they are inefficient for compressing sing le images. As an alternative, we propose a novel method, Relative Entropy Coding (REC), that can directly encode the latent representation with codelength close to the relative entropy for single images, supported by our empirical results o btained on the Cifar10, ImageNet32 and Kodak datasets. Moreover, unlike previous bits-back methods, REC is immediately applicable to lossy compression, where it is competitive with the state-of-the-art on the Kodak dataset.
**********

Improved Guarantees for k-means++ and k-means++ Parallel
Konstantin Makarychev, Aravind Reddy, Liren Shan
In this paper, we study k-means++ and k-means||, the two most popular algorithms for the classic k-means clustering problem. We provide novel analyses and show improved approximation and bi-criteria approximation guarantees for k-means++ an d k-means||. Our results give a better theoretical justification for why these a lgorithms perform extremely well in practice.
**********

Sparse Spectrum Warped Input Measures for Nonstationary Kernel Learning
Anthony Tompkins, Rafael Oliveira, Fabio T. Ramos
We establish a general form of explicit, input-dependent, measure-valued warping s for learning nonstationary kernels. While stationary kernels are uniquitous an d simple to use, they struggle to adapt to functions that vary in smoothness wit h respect to the input. The proposed learning algorithm warps inputs as conditio nal Gaussian measures that control the smoothness of a standard stationary kerne l. This construction allows us to capture non-stationary patterns in the data an d provides intuitive inductive bias. The resulting method is based on sparse spe ctrum Gaussian processes, enabling closed-form solutions, and is extensible to a

stacked construction to capture more complex patterns. The method is extensively validated alongside related algorithms on synthetic and real world datasets. We demonstrate a remarkable efficiency in the number of parameters of the warping functions in learning problems with both small and large data regimes.
**********

An Efficient Adversarial Attack for Tree Ensembles
Chong Zhang, Huan Zhang, 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.
**********

Learning Continuous System Dynamics from Irregularly-Sampled Partial Observations
Zijie Huang, Yizhou Sun, Wei Wang
Many real-world systems, such as moving planets, can be considered as multi-agent dynamic systems, where objects interact with each other and co-evolve along with the time. Such dynamics is usually difficult to capture, and understanding and predicting the dynamics based on observed trajectories of objects become a critical research problem in many domains. Most existing algorithms, however, assume the observations are regularly sampled and all the objects can be fully observed at each sampling time, which is impractical for many applications. In this paper, we pro-pose to learn system dynamics from irregularly-sampled and partial observations with underlying graph structure for the first time. To tackle the above challenge, we present LG-ODE, a latent ordinary differential equation generative model for modeling multi-agent dynamic system with known graph structure. It can simultaneously learn the embedding of high dimensional trajectories and infer continuous latent system dynamics. Our model employs a novel encoder parameterized by a graph neural network that can infer initial states in an unsupervised way from irregularly-sampled partial observations of structural objects and utilizes neuralODE to infer arbitrarily complex continuous-time latent dynamics. Experiments on motion capture, spring system, and charged particle datasets demonstrate the effectiveness of our approach.
**********

Online Bayesian Persuasion
Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Nicola Gatti
In Bayesian persuasion, an informed sender has to design a signaling scheme that discloses the right amount of information so as to influence the behavior of a self-interested receiver. This kind of strategic interaction is ubiquitous in real economic scenarios. However, the original model by Kamenica and Gentzkow makes some stringent assumptions which limit its applicability in practice. One of the most limiting assumptions is arguably that, in order to compute an optimal signaling scheme, the sender is usually required to know the receiver's utility function. In this paper, we relax this assumption through an online learning framework in which the sender faces a receiver with unknown type. At each round, the receiver's type is chosen adversarially from a finite set of possible types. We are interested in no-regret algorithms prescribing a signaling scheme at each round of the repeated interaction with performances close to that of the best-in-hindsight signaling scheme. First, we prove a hardness result on the per-iteration running time required to achieve the no-regret property. Then, we provide algorithms for the full and partial information model which exhibit regret sublinear in the number of rounds and polynomial in the parameters of the game.
**********

Robust Pre-Training by Adversarial Contrastive Learning
Ziyu Jiang, Tianlong Chen, Ting Chen, Zhangyang Wang
Recent work has shown that, when integrated with adversarial training, self-supervised pre-training can lead to state-of-the-art robustness In this work, we improve robustness-aware self-supervised pre-training by learning representations that are consistent under both data augmentations and adversarial perturbations. Our approach leverages a recent contrastive learning framework, which learns rep

resentations by maximizing feature consistency under differently augmented views . This fits particularly well with the goal of adversarial robustness, as one ca use of adversarial fragility is the lack of feature invariance, i.e., small inpu t perturbations can result in undesirable large changes in features or even pred icted labels. We explore various options to formulate the contrastive task, and demonstrate that by injecting adversarial perturbations, contrastive pre-trainin g can lead to models that are both label-efficient and robust. We empirically ev aluate the proposed Adversarial Contrastive Learning (ACL) and show it can consi stently outperform existing methods. For example on the CIFAR-10 dataset, ACL ou tperforms the previous state-of-the-art unsupervised robust pre-training approac h by 2.99% on robust accuracy and 2.14% on standard accuracy. We further demonst rate that ACL pre-training can improve semi-supervised adversarial training, eve n when only a few labeled examples are available. Our codes and pre-trained mode ls have been released at: https://github.com/VITA-Group/Adversarial-Contrastive-Learning.

**********

Random Walk Graph Neural Networks

Giannis Nikolentzos, Michalis Vazirgiannis

In recent years, graph neural networks (GNNs) have become the de facto tool for performing machine learning tasks on graphs. Most GNNs belong to the family of m essage passing neural networks (MPNNs). These models employ an iterative neighbo rhood aggregation scheme to update vertex representations. Then, to compute vect or representations of graphs, they aggregate the representations of the vertices using some permutation invariant function. One would expect the hidden layers o f a GNN to be composed of parameters that take the form of graphs. However, this is not the case for MPNNs since their update procedure is parameterized by full y-connected layers. In this paper, we propose a more intuitive and transparent a rchitecture for graph-structured data, so-called Random Walk Graph Neural Networ k (RWNN). The first layer of the model consists of a number of trainable ``hidde n graphs'' which are compared against the input graphs using a random walk kerne l to produce graph representations. These representations are then passed on to a fully-connected neural network which produces the output. The employed random walk kernel is differentiable, and therefore, the proposed model is end-to-end t rainable. We demonstrate the model's transparency on synthetic datasets. Further more, we empirically evaluate the model on graph classification datasets and sho w that it achieves competitive performance.

**********

Explore Aggressively, Update Conservatively: Stochastic Extragradient Methods wi th Variable Stepsize Scaling

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, Panayotis Mertikopoulos

Owing to their stability and convergence speed, extragradient methods have becom e a staple for solving large-scale saddle-point problems in machine learning. Th e basic premise of these algorithms is the use of an extrapolation step before p erforming an update; thanks to this exploration step, extra-gradient methods ove rcome many of the non-convergence issues that plague gradient descent/ascent sch emes. On the other hand, as we show in this paper, running vanilla extragradient with stochastic gradients may jeopardize its convergence, even in simple biline ar models. To overcome this failure, we investigate a double stepsize extragradi ent algorithm where the exploration step evolves at a more aggressive time-scale compared to the update step. We show that this modification allows the method t o converge even with stochastic gradients, and we derive sharp convergence rates under an error bound condition.

**********

Fast and Accurate $k$-means++ via Rejection Sampling

Vincent Cohen-Addad, Silvio Lattanzi, Ashkan Norouzi-Fard, Christian Sohler, Ola Svensson

Requests for name changes in the electronic proceedings will be accepted with 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.

**********
Variational Amodal Object Completion
Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, Sanja Fidler
In images of complex scenes, objects are often occluding each other which makes perception tasks such as object detection and tracking, or robotic control tasks such as planning, challenging. To facilitate downstream tasks, it is thus important to reason about the full extent of objects, i.e., seeing behind occlusion, typically referred to as amodal instance completion. In this paper, we propose a variational generative framework for amodal completion, referred to as AMODAL-VAE, which does not require any amodal labels at training time, as it is able to utilize widely available object instance masks. We showcase our approach on the downstream task of scene editing where the user is presented with interactive tools to complete and erase objects in photographs. Experiments on complex street scenes demonstrate state-of-the-art performance in amodal mask completion and showcase high-quality scene editing results. Interestingly, a user study shows that humans prefer object completions inferred by our model to the human-labeled ones.
**********
When Counterpoint Meets Chinese Folk Melodies
Nan Jiang, Sheng Jin, Zhiyao Duan, Changshui Zhang
Counterpoint is an important concept in Western music theory. In the past century, there have been significant interests in incorporating counterpoint into Chinese folk music composition. In this paper, we propose a reinforcement learning-based system, named FolkDuet, towards the online countermelody generation for Chinese folk melodies. With no existing data of Chinese folk duets, FolkDuet employs two reward models based on out-of-domain data, i.e. Bach chorales, and monophonic Chinese folk melodies. An interaction reward model is trained on the duets formed from outer parts of Bach chorales to model counterpoint interaction, while a style reward model is trained on monophonic melodies of Chinese folk songs to model melodic patterns. With both rewards, the generator of FolkDuet is trained to generate countermelodies while maintaining the Chinese folk style. The entire generation process is performed in an online fashion, allowing real-time interactive human-machine duet improvisation. Experiments show that the proposed algorithm achieves better subjective and objective results than the baselines.
**********
Sub-linear Regret Bounds for Bayesian Optimisation in Unknown Search Spaces
Hung Tran-The, Sunil Gupta, Santu Rana, Huong Ha, Svetha Venkatesh
Bayesian optimisation is a popular method for efficient optimisation of expensive black-box functions. Traditionally, BO assumes that the search space is known. However, in many problems, this assumption does not hold. To this end, we propose a novel BO algorithm which expands (and shifts) the search space over iterations based on controlling the expansion rate thought a \emph{hyperharmonic series}. Further, we propose another variant of our algorithm that scales to high dimensions. We show theoretically that for both our algorithms, the cumulative regret grows at sub-linear rates. Our experiments with synthetic and real-world optimisation tasks demonstrate the superiority of our algorithms over the current state-of-the-art methods for Bayesian optimisation in unknown search space.
**********
Universal Domain Adaptation through Self Supervision
Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Kate Saenko
Unsupervised domain adaptation methods traditionally assume that all source categories are present in the target domain. In practice, little may be known about the category overlap between the two domains. While some methods address target settings with either partial or open-set categories, they assume that the particular setting is known a priori. We propose a more universally applicable domain adaptation approach that can handle arbitrary category shift, called Domain Adaptative Neighborhood Clustering via Entropy optimization (DANCE). Our approach combines two novel ideas: First, as we cannot fully rely on source categories to learn features discriminative for the target, we propose a novel neighborhood clustering technique to learn the structure of the target domain in a self-supervis

ed way. Second, we use entropy-based feature alignment and rejection to align ta
rget features with the source, or reject them as unknown categories based on the
ir entropy.
We show through extensive experiments that DANCE outperforms baselines across op
en-set, open-partial and partial domain adaptation settings.
**********

Patch2Self: Denoising Diffusion MRI with Self-Supervised Learning■
Shreyas Fadnavis, Joshua Batson, Eleftherios Garyfallidis
Diffusion-weighted magnetic resonance imaging (DWI) is the only non-invasive met
hod for quantifying microstructure and reconstructing white-matter pathways in t
he living human brain. Fluctuations from multiple sources create significant noi
se in DWI data which must be suppressed before subsequent microstructure analysi
s. We introduce a self-supervised learning method for denoising DWI data, Patch2
Self, which uses the entire volume to learn a full-rank locally linear denoiser
for that volume. By taking advantage of the oversampled q-space of DWI data, Pat
ch2Self can separate structure from noise without requiring an explicit model fo
r either. We demonstrate the effectiveness of Patch2Self via quantitative and qu
alitative improvements in microstructure modeling, tracking (via fiber bundle co
herency) and model estimation relative to other unsupervised methods on real and
 simulated data.
**********

Stochastic Normalization
Zhi Kou, Kaichao You, Mingsheng Long, Jianmin Wang
Fine-tuning pre-trained deep networks on a small dataset is an important compone
nt in the deep learning pipeline. A critical problem in fine-tuning is how to av
oid over-fitting when data are limited. Existing efforts work from two aspects:
(1) impose regularization on parameters or features; (2) transfer prior knowledg
e to fine-tuning by reusing pre-trained parameters. In this paper, we take an al
ternative approach by refactoring the widely used Batch Normalization (BN) modul
e to mitigate over-fitting. We propose a two-branch design with one branch norma
lized by mini-batch statistics and the other branch normalized by moving statist
ics. During training, two branches are stochastically selected to avoid over-dep
ending on some sample statistics, resulting in a strong regularization effect, w
hich we interpret as ``architecture regularization.'' The resulting method is du
bbed stochastic normalization (\textbf{StochNorm}). With the two-branch architec
ture, it naturally incorporates pre-trained moving statistics in BN layers durin
g fine-tuning, exploiting more prior knowledge of pre-trained networks. Extensiv
e empirical experiments show that StochNorm is a powerful tool to avoid over-fit
ting in fine-tuning with small datasets. Besides, StochNorm is readily pluggable
 in modern CNN backbones. It is complementary to other fine-tuning methods and c
an work together to achieve stronger regularization effect.
**********

Constrained episodic reinforcement learning in concave-convex and knapsack setti
ngs
Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowit
z, Aleksandrs Slivkins, Wen Sun
We propose an algorithm for tabular episodic reinforcement learning with constra
ints. We provide a modular analysis with strong theoretical guarantees for setti
ngs with concave rewards and convex constraints, and for settings with hard cons
traints (knapsacks). Most of the previous work in constrained reinforcement lear
ning is limited to linear constraints, and the remaining work focuses on either
the feasibility question or settings with a single episode. Our experiments demo
nstrate that the proposed algorithm significantly outperforms these approaches i
n existing constrained episodic environments.
**********

On Learning Ising Models under Huber's Contamination Model
Adarsh Prasad, Vishwak Srinivasan, Sivaraman Balakrishnan, Pradeep Ravikumar
We study the problem of learning Ising models in a setting where some of the sam
ples from the underlying distribution can be arbitrarily corrupted. In such a se
tup, we aim to design statistically optimal estimators in a

high-dimensional scaling in which the number of nodes p, the number of edges k a
nd the maximal node degree d are allowed to increase to infinity as a function o
f the sample size n. Our analysis is based on exploiting moments of the underlyi
ng distribution, coupled with novel reductions to univariate estimation. Our pro
posed estimators achieve an optimal dimension independent dependence on the frac
tion of corrupted data in the contaminated setting, while also simultaneously ac
hieving high-probability error guarantees with optimal sample-complexity. We cor
roborate our theoretical results by simulations.
**********

Cross-validation Confidence Intervals for Test Error
Pierre Bayle, Alexandre Bayle, Lucas Janson, Lester Mackey
This work develops central limit theorems for cross-validation and consistent es
timators of the asymptotic variance under weak stability conditions on the learn
ing algorithm. Together, these results provide practical, asymptotically-exact c
onfidence intervals for k-fold test error and valid, powerful hypothesis tests o
f whether one learning algorithm has smaller k-fold test error than another. The
se results are also the first of their kind for the popular choice of leave-one-
out cross-validation. In our experiments with diverse learning algorithms, the r
esulting intervals and tests outperform the most popular alternative methods fro
m the literature.
**********

DeepSVG: A Hierarchical Generative Network for Vector Graphics Animation
Alexandre Carlier, Martin Danelljan, Alexandre Alahi, Radu Timofte
Scalable Vector Graphics (SVG) are ubiquitous in modern 2D interfaces due to the
ir ability to scale to different resolutions. However, despite the success of de
ep learning-based models applied to rasterized images, the problem of vector gra
phics representation learning and generation remains largely unexplored. In this
 work, we propose a novel hierarchical generative network, called DeepSVG, for c
omplex SVG icons generation and interpolation. Our architecture effectively dise
ntangles high-level shapes from the low-level commands that encode the shape its
elf. The network directly predicts a set of shapes in a non-autoregressive fashi
on. We introduce the task of complex SVG icons generation by releasing a new lar
ge-scale dataset along with an open-source library for SVG manipulation. We demo
nstrate that our network learns to accurately reconstruct diverse vector graphic
s, and can serve as a powerful animation tool by performing interpolations and o
ther latent space operations. Our code is available at https://github.com/alexan
dre01/deepsvg.
**********

Bayesian Attention Modules
Xinjie Fan, Shujian Zhang, Bo Chen, Mingyuan Zhou
Attention modules, as simple and effective tools, have not only enabled deep neu
ral networks to achieve state-of-the-art results in many domains, but also enhan
ced their interpretability. Most current models use deterministic attention modu
les due to their simplicity and ease of optimization. Stochastic counterparts, o
n the other hand, are less popular despite their potential benefits. The main re
ason is that stochastic attention often introduces optimization issues or requir
es significant model changes. In this paper, we propose a scalable stochastic ve
rsion of attention that is easy to implement and optimize. We construct simplex-
constrained attention distributions by normalizing reparameterizable distributio
ns, making the training process differentiable. We learn their parameters in a B
ayesian framework where a data-dependent prior is introduced for regularization.
 We apply the proposed stochastic attention modules to various attention-based m
odels, with applications to graph node classification, visual question answering
, image captioning, machine translation, and language understanding. Our experim
ents show the proposed method brings consistent improvements over the correspond
ing baselines.
**********

Robustness Analysis of Non-Convex Stochastic Gradient Descent using Biased Expec
tations
Kevin Scaman, Cedric Malherbe

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

SoftFlow: Probabilistic Framework for Normalizing Flow on Manifolds
Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, Nam Soo Kim
Flow-based generative models are composed of invertible transformations between two random variables of the same dimension. Therefore, flow-based models cannot be adequately trained if the dimension of the data distribution does not match that of the underlying target distribution. In this paper, we propose SoftFlow, a probabilistic framework for training normalizing flows on manifolds. To sidestep the dimension mismatch problem, SoftFlow estimates a conditional distribution of the perturbed input data instead of learning the data distribution directly. We experimentally show that SoftFlow can capture the innate structure of the manifold data and generate high-quality samples unlike the conventional flow-based models. Furthermore, we apply the proposed framework to 3D point clouds to alleviate the difficulty of forming thin structures for flow-based models. The proposed model for 3D point clouds, namely SoftPointFlow, can estimate the distribution of various shapes more accurately and achieves state-of-the-art performance in point cloud generation.
**********

A meta-learning approach to (re)discover plasticity rules that carve a desired function into a neural network
Basile Confavreux, Friedemann Zenke, Everton Agnes, Timothy Lillicrap, Tim Vogels
The search for biologically faithful synaptic plasticity rules has resulted in a large body of models. They are usually inspired by -- and fitted to -- experimental data, but they rarely produce neural dynamics that serve complex functions. These failures suggest that current plasticity models are still under-constrained by existing data. Here, we present an alternative approach that uses meta-learning to discover plausible synaptic plasticity rules. Instead of experimental data, the rules are constrained by the functions they implement and the structure they are meant to produce. Briefly, we parameterize synaptic plasticity rules by a Volterra expansion and then use supervised learning methods (gradient descent or evolutionary strategies) to minimize a problem-dependent loss function that quantifies how effectively a candidate plasticity rule transforms an initially random network into one with the desired function. We first validate our approach by re-discovering previously described plasticity rules, starting at the single-neuron level and ``Oja’s rule'', a simple Hebbian plasticity rule that captures the direction of most variability of inputs to a neuron (i.e., the first principal component). We expand the problem to the network level and ask the framework to find Oja’s rule together with an anti-Hebbian rule such that an initially random two-layer firing-rate network will recover several principal components of the input space after learning. Next, we move to networks of integrate-and-fire neurons with plastic inhibitory afferents. We train for rules that achieve a target firing rate by countering tuned excitation. Our algorithm discovers a specific subset of the manifold of rules that can solve this task. Our work is a proof of principle of an automated and unbiased approach to unveil synaptic plasticity rules that obey biological constraints and can solve complex functions.
**********

Greedy Optimization Provably Wins the Lottery: Logarithmic Number of Winning Tickets is Enough
Mao Ye, Lemeng Wu, Qiang Liu
Despite the great success of deep learning, recent works show that large deep neural networks are often highly redundant and can be significantly reduced in size. However, the theoretical question of how much we can prune a neural network given a specified tolerance of accuracy drop is still open. This paper provides one answer to this question by proposing a greedy optimization based pruning method. The proposed method has the guarantee that the discrepancy between the prune

d network and the original network decays with exponentially fast rate w.r.t. th
e size of the pruned network, under weak assumptions that apply for most practic
al settings. Empirically, our method improves prior arts on pruning various netw
ork architectures including ResNet, MobilenetV2/V3 on ImageNet.
**********

Path Integral Based Convolution and Pooling for Graph Neural Networks
Zheng Ma, Junyu Xuan, Yu Guang Wang, Ming Li, Pietro Liò
Graph neural networks (GNNs) extends the functionality of traditional neural net
works to graph-structured data. Similar to CNNs, an optimized design of graph co
nvolution and pooling is key to success. Borrowing ideas from physics, we propos
e a path integral based graph neural networks (PAN) for classification and regre
ssion tasks on graphs. Specifically, we consider a convolution operation that in
volves every path linking the message sender and receiver with learnable weights
 depending on the path length, which corresponds to the maximal entropy random w
alk. It generalizes the graph Laplacian to a new transition matrix we call \emph
{maximal entropy transition} (MET) matrix derived from a path integral formalism
. Importantly, the diagonal entries of the MET matrix are directly related to th
e subgraph centrality, thus lead to a natural and adaptive pooling mechanism. PA
N provides a versatile framework that can be tailored for different graph data w
ith varying sizes and structures. We can view most existing GNN architectures as
 special cases of PAN. Experimental results show that PAN achieves state-of-the-
art performance on various graph classification/regression tasks, including a ne
w benchmark dataset from statistical mechanics we propose to boost applications
of GNN in physical sciences.
**********

Estimating the Effects of Continuous-valued Interventions using Generative Adver
sarial Networks
Ioana Bica, James Jordon, Mihaela van der Schaar
While much attention has been given to the problem of estimating the effect of d
iscrete interventions from observational data, relatively little work has been d
one in the setting of continuous-valued interventions, such as treatments associ
ated with a dosage parameter. In this paper, we tackle this problem by building
on a modification of the generative adversarial networks (GANs) framework. Our m
odel, SCIGAN, is flexible and capable of simultaneously estimating counterfactua
l outcomes for several different continuous interventions. The key idea is to us
e a significantly modified GAN model to learn to generate counterfactual outcome
s, which can then be used to learn an inference model, using standard supervised
 methods, capable of estimating these counterfactuals for a new sample. To addre
ss the challenges presented by shifting to continuous interventions, we propose
a novel architecture for our discriminator - we build a hierarchical discriminat
or that leverages the structure of the continuous intervention setting. Moreover
, we provide theoretical results to support our use of the GAN framework and of
the hierarchical discriminator. In the experiments section, we introduce a new s
emi-synthetic data simulation for use in the continuous intervention setting and
 demonstrate improvements over the existing benchmark models.
**********

Latent Dynamic Factor Analysis of High-Dimensional Neural Recordings
Heejong Bong, Zongge Liu, Zhao Ren, Matthew Smith, Valerie Ventura, Robert E. Ka
ss
High-dimensional neural recordings across multiple brain regions can be used to
establish functional connectivity with good spatial and temporal resolution. We
designed and implemented a novel method, Latent Dynamic Factor Analysis of High-
dimensional time series (LDFA-H), which combines (a) a new approach to estimatin
g the covariance structure among high-dimensional time series (for the observed
variables) and (b) a new extension of probabilistic CCA to dynamic time series (
for the latent variables). Our interest is in the cross-correlations among the l
atent variables which, in neural recordings, may capture the flow of information
 from one brain region to another. Simulations show that LDFA-H outperforms exis
ting methods in the sense that it captures target factors even when within-regio
n correlation due to noise dominates cross-region correlation. We applied our me

thod to local field potential (LFP) recordings from 192 electrodes in Prefrontal Cortex (PFC) and visual area V4 during a memory-guided saccade task. The results capture time-varying lead-lag dependencies between PFC and V4, and display the associated spatial distribution of the signals.
**********

Conditioning and Processing: Techniques to Improve Information-Theoretic Generalization Bounds
Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, Mahdieh Soleymani
Obtaining generalization bounds for learning algorithms is one of the main subjects studied in theoretical machine learning. In recent years, information-theoretic bounds on generalization have gained the attention of researchers. This approach provides an insight into learning algorithms by considering the mutual information between the model and the training set. In this paper, a probabilistic graphical representation of this approach is adopted and two general techniques to improve the bounds are introduced, namely conditioning and processing. In conditioning, a random variable in the graph is considered as given, while in processing a random variable is substituted with one of its children. These techniques can be used to improve the bounds by either sharpening them or increasing their applicability. It is demonstrated that the proposed framework provides a simple and unified way to explain a variety of recent tightening results. New improved bounds derived by utilizing these techniques are also proposed.
**********

Bongard-LOGO: A New Benchmark for Human-Level Concept Learning and Reasoning
Weili Nie, Zhiding Yu, Lei Mao, Ankit B. Patel, Yuke Zhu, Anima Anandkumar
Humans have an inherent ability to learn novel concepts from only a few samples and generalize these concepts to different situations. Even though today's machine learning models excel with a plethora of training data on standard recognition tasks, a considerable gap exists between machine-level pattern recognition and human-level concept learning. To narrow this gap, the Bongard Problems (BPs) were introduced as an inspirational challenge for visual cognition in intelligent systems. Albeit new advances in representation learning and learning to learn, BPs remain a daunting challenge for modern AI. Inspired by the original one hundred BPs, we propose a new benchmark Bongard-LOGO for human-level concept learning and reasoning. We develop a program-guided generation technique to produce a large set of human-interpretable visual cognition problems in action-oriented LOGO language. Our benchmark captures three core properties of human cognition: 1) context-dependent perception, in which the same object may have disparate interpretations given different contexts; 2) analogy-making perception, in which some meaningful concepts are traded off for other meaningful concepts; and 3) perception with a few samples but infinite vocabulary. In experiments, we show that the state-of-the-art deep learning methods perform substantially worse than human subjects, implying that they fail to capture core human cognition properties. Finally, we discuss  research directions towards a general architecture for visual reasoning to tackle this benchmark.
**********

GAN Memory with No Forgetting
Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, Lawrence Carin
As a fundamental issue in lifelong learning, catastrophic forgetting is directly caused by inaccessible historical data; accordingly, if the data (information) were memorized perfectly, no forgetting should be expected. Motivated by that, we propose a GAN memory for lifelong learning, which is capable of remembering a stream of datasets via generative processes, with \emph{no} forgetting. Our GAN memory is based on recognizing that one can modulate the ``style'' of a GAN model to form perceptually-distant targeted generation. Accordingly, we propose to do sequential style modulations atop a well-behaved base GAN model, to form sequential targeted generative models, while simultaneously benefiting from the transferred base knowledge. The GAN memory -- that is motivated by lifelong learning -- is therefore itself manifested by a form of lifelong learning, via forward transfer and modulation of information from prior tasks. Experiments demonstrate the superiority of our method over existing approaches and its effectiveness in a

lleviating catastrophic forgetting for lifelong classification problems. Code is available at \url{https://github.com/MiaoyunZhao/GANmemory_LifelongLearning}.
**********
Deep Reinforcement Learning with Stacked Hierarchical Attention for Text-based Games
Yunqiu Xu, Meng Fang, Ling Chen, Yali Du, Joey Tianyi Zhou, Chengqi Zhang
We study reinforcement learning (RL) for text-based games, which are interactive simulations in the context of natural language. While different methods have been developed to represent the environment information and language actions, existing RL agents are not empowered with any reasoning capabilities to deal with textual games. In this work, we aim to conduct explicit reasoning with knowledge graphs for decision making, so that the actions of an agent are generated and supported by an interpretable inference procedure. We propose a stacked hierarchical attention mechanism to construct an explicit representation of the reasoning process by exploiting the structure of the knowledge graph. We extensively evaluate our method on a number of man-made benchmark games, and the experimental results demonstrate that our method performs better than existing text-based agents.
**********
Gaussian Gated Linear Networks
David Budden, Adam Marblestone, Eren Sezener, Tor Lattimore, Gregory Wayne, Joel Veness
We propose the Gaussian Gated Linear Network (G-GLN), an extension to the recently proposed GLN family of deep neural networks. Instead of using backpropagation to learn features, GLNs have a distributed and local credit assignment mechanism based on optimizing a convex objective. This gives rise to many desirable properties including universality, data-efficient online learning, trivial interpretability and robustness to catastrophic forgetting. We extend the GLN framework from classification to multiple regression and density modelling by generalizing geometric mixing to a product of Gaussian densities. The G-GLN achieves competitive or state-of-the-art performance on several univariate and multivariate regression benchmarks, and we demonstrate its applicability to practical tasks including online contextual bandits and density estimation via denoising.
**********
Node Classification on Graphs with Few-Shot Novel Labels via Meta Transformed Network Embedding
Lin Lan, Pinghui Wang, Xuefeng Du, Kaikai Song, Jing Tao, Xiaohong Guan
We study the problem of node classification on graphs with few-shot novel labels, which has two distinctive properties: (1) There are novel labels to emerge in the graph; (2) The novel labels have only a few representative nodes for training a classifier. The study of this problem is instructive and corresponds to many applications such as recommendations for newly formed groups with only a few users in online social networks. To cope with this problem, we propose a novel Meta Transformed Network Embedding framework (MetaTNE), which consists of three modules: (1) A \emph{structural module} provides each node a latent representation according to the graph structure. (2) A \emph{meta-learning module} captures the relationships between the graph structure and the node labels as prior knowledge in a meta-learning manner. Additionally, we introduce an \emph{embedding transformation function} that remedies the deficiency of the straightforward use of meta-learning. Inherently, the meta-learned prior knowledge can be used to facilitate the learning of few-shot novel labels. (3) An \emph{optimization module} employs a simple yet effective scheduling strategy to train the above two modules with a balance between graph structure learning and meta-learning. Experiments on four real-world datasets show that MetaTNE brings a huge improvement over the state-of-the-art methods.
**********
Online Fast Adaptation and Knowledge Accumulation (OSAKA): a New Approach to Continual Learning
Massimo Caccia, Pau Rodriguez, Oleksiy Ostapenko, Fabrice Normandin, Min Lin, Lucas Page-Caccia, Issam Hadj Laradji, Irina Rish, Alexandre Lacoste, David Vázquez, Laurent Charlin

Continual learning agents experience a stream of (related) tasks. The main challenge is that the agent must not forget previous tasks and also adapt to novel tasks in the stream. We are interested in the intersection of two recent continual-learning scenarios. In meta-continual learning, the model is pre-trained using meta-learning to minimize catastrophic forgetting of previous tasks. In continual-meta learning, the aim is to train agents for faster remembering of previous tasks through adaptation. In their original formulations, both methods have limitations. We stand on their shoulders to propose a more general scenario, OSAKA, where an agent must quickly solve new (out-of-distribution) tasks, while also requiring fast remembering. We show that current continual learning, meta-learning, meta-continual learning, and continual-meta learning techniques fail in this new scenario.
We propose Continual-MAML, an online extension of the popular MAML algorithm as a strong baseline for this scenario. We show in an empirical study that Continual-MAML is better suited to the new scenario than the aforementioned methodologies including standard continual learning and meta-learning approaches.
**********
Convex optimization based on global lower second-order models
Nikita Doikov, Yurii Nesterov
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Simultaneously Learning Stochastic and Adversarial Episodic MDPs with Known Transition
Tiancheng Jin, Haipeng Luo
This work studies the problem of learning episodic Markov Decision Processes with known transition and bandit feedback. We develop the first algorithm with a ``best-of-both-worlds'' guarantee: it achieves O(log T) regret when the losses are stochastic, and simultaneously enjoys worst-case robustness with \tilde{O}(\sqrt{T}) regret even when the losses are adversarial, where T is the number of episodes. More generally, it achieves \tilde{O}(\sqrt{C}) regret in an intermediate setting where the losses are corrupted by a total amount of C.
Our algorithm is based on the Follow-the-Regularized-Leader method from Zimin and Neu (2013), with a novel hybrid regularizer inspired by recent works of Zimmert et al. (2019a, 2019b) for the special case of multi-armed bandits. Crucially, our regularizer admits a non-diagonal Hessian with a highly complicated inverse. Analyzing such a regularizer and deriving a particular self-bounding regret guarantee is our key technical contribution and might be of independent interest.
**********
Relative gradient optimization of the Jacobian term in unsupervised deep learning
Luigi Gresele, Giancarlo Fissore, Adrián Javaloy, Bernhard Schölkopf, Aapo Hyvarinen
Learning expressive probabilistic models correctly describing the data is a ubiquitous problem in machine learning. A popular approach for solving it is mapping the observations into a representation space with a simple joint distribution, which can typically be written as a product of its marginals — thus drawing a connection with the field of nonlinear independent component analysis. Deep density models have been widely used for this task, but their maximum likelihood based training requires estimating the log-determinant of the Jacobian and is computationally expensive, thus imposing a trade-off between computation and expressive power. In this work, we propose a new approach for exact training of such neural networks. Based on relative gradients, we exploit the matrix structure of neural network parameters to compute updates efficiently even in high-dimensional spaces; the computational cost of the training is quadratic in the input size, in contrast with the cubic scaling of naive approaches. This allows fast training with objective functions involving the log-determinant of the Jacobian, without imposing constraints on its structure, in stark contrast to autoregressive normal

izing flows.
**********
Self-Supervised Visual Representation Learning from Hierarchical Grouping
Xiao Zhang, Michael Maire
We create a framework for bootstrapping visual representation learning from a pr imitive visual grouping capability. We operationalize grouping via a contour det ector that partitions an image into regions, followed by merging of those region s into a tree hierarchy. A small supervised dataset suffices for training this g rouping primitive. Across a large unlabeled dataset, we apply this learned primi tive to automatically predict hierarchical region structure. These predictions s erve as guidance for self-supervised contrastive feature learning: we task a dee p network with producing per-pixel embeddings whose pairwise distances respect t he region hierarchy. Experiments demonstrate that our approach can serve as stat e-of-the-art generic pre-training, benefiting downstream tasks. We additionally explore applications to semantic region search and video-based object instance t racking.
**********
Optimal Variance Control of the Score-Function Gradient Estimator for Importance -Weighted Bounds
Valentin Liévin, Andrea Dittadi, Anders Christensen, Ole Winther
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
Explicit Regularisation in Gaussian Noise Injections
Alexander Camuto, Matthew Willetts, Umut Simsekli, Stephen J. Roberts, Chris C. Holmes
We study the regularisation induced in neural networks by Gaussian noise injecti ons (GNIs). Though such injections have been extensively studied when applied to data, there have been few studies on understanding the regularising effect they induce when applied to network activations. Here we derive the explicit regular iser of GNIs, obtained by marginalising out the injected noise, and show that it penalises functions with high-frequency components in the Fourier domain; parti cularly in layers closer to a neural network's output. We show analytically and empirically that such regularisation produces calibrated classifiers with large classification margins.
**********
Numerically Solving Parametric Families of High-Dimensional Kolmogorov Partial D ifferential Equations via Deep Learning
Julius Berner, Markus Dablander, Philipp Grohs
We present a deep learning algorithm for the numerical solution of parametric fa milies of high-dimensional linear Kolmogorov partial differential equations (PDE s). Our method is based on reformulating the numerical approximation of a whole family of Kolmogorov PDEs as a single statistical learning problem using the Fey nman-Kac formula. Successful numerical experiments are presented, which empirica lly confirm the functionality and efficiency of our proposed algorithm in the ca se of heat equations and Black-Scholes option pricing models parametrized by aff ine-linear coefficient functions. We show that a single deep neural network trai ned on simulated data is capable of learning the solution functions of an entire family of PDEs on a full space-time region. Most notably, our numerical observa tions and theoretical results also demonstrate that the proposed method does not suffer from the curse of dimensionality, distinguishing it from almost all stan dard numerical methods for PDEs.
**********
Finite-Time Analysis for Double Q-learning
Huaqing Xiong, Lin Zhao, Yingbin Liang, Wei 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.
**********
Learning to Detect Objects with a 1 Megapixel Event Camera
Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, Amos Sironi
Event cameras encode visual information with high temporal precision,
low data-rate, and high-dynamic range.
Thanks to these characteristics, event cameras are particularly suited for scenarios
with high motion, challenging lighting conditions and requiring low latency.
However, due to the novelty of the field, the performance of event-based systems
on many vision tasks is still lower compared to conventional frame-based solutions.
The main reasons for this performance gap are: the lower spatial resolution of event sensors,
compared to frame cameras; the lack of large-scale training datasets;
the absence of well established deep learning architectures for event-based processing.
In this paper, we address all these problems in the context of an event-based object detection task.
First, we publicly release the first high-resolution large-scale dataset for object detection.
The dataset contains more than 14 hours recordings of a 1 megapixel event camera,
in automotive scenarios, together with 25M bounding boxes of
cars, pedestrians, and two-wheelers, labeled at high frequency.
Second, we introduce a novel recurrent architecture for event-based detection
and a temporal consistency loss for better-behaved training.
The ability to compactly represent the sequence of events into the internal memory
of the model is essential to achieve high accuracy. Our model outperforms by a
large margin feed-forward event-based architectures.
Moreover, our method does not require any reconstruction
of intensity images from events, showing that training directly from raw events is possible,
more efficient, and more accurate than passing through an intermediate intensity image.
Experiments on the dataset introduced in this work, for which events
and gray level images are available, show performance on par with that of
highly tuned and studied frame-based detectors.
**********
End-to-End Learning and Intervention in Games
Jiayang Li, Jing Yu, Yu Nie, Zhaoran Wang
In a social system, the self-interest of agents can be detrimental to the collective good, sometimes leading to social dilemmas. To resolve such a conflict, a central designer may intervene by either redesigning the system or incentivizing the agents to change their behaviors. To be effective, the designer must anticipate how the agents react to the intervention, which is dictated by their often unknown payoff functions. Therefore, learning about the agents is a prerequisite for intervention. In this paper, we provide a unified framework for learning and intervention in games. We cast the equilibria of games as individual layers and integrate them into an end-to-end optimization framework. To enable the backward propagation through the equilibria of games, we propose two approaches, respectively based on explicit and implicit differentiation. Specifically, we cast the equilibria as the solutions to variational inequalities (VIs). The explicit approach unrolls the projection method for solving VIs, while the implicit approach exploits the sensitivity of the solutions to VIs. At the core of both approaches is the differentiation through a projection operator. Moreover, we establish the correctness of both approaches and identify the conditions under which one approach is more desirable than the other. The analytical results are validated

using several real-world problems.
**********
Least Squares Regression with Markovian Data: Fundamental Limits and Algorithms
Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, 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.
**********
Predictive coding in balanced neural networks with noise, chaos and delays
Jonathan Kadmon, Jonathan Timcheck, Surya Ganguli
Biological neural networks face a formidable task: performing reliable computations in the face of intrinsic stochasticity in individual neurons, imprecisely specified synaptic connectivity, and nonnegligible delays in synaptic transmission. A common approach to combatting such biological heterogeneity involves averaging over large redundant networks of N neurons resulting in coding errors that decrease classically as the square root of N. Recent work demonstrated a novel mechanism whereby recurrent spiking networks could efficiently encode dynamic stimuli achieving a superclassical scaling in which coding errors decrease as 1/N. This specific mechanism involved two key ideas: predictive coding, and a tight balance, or cancellation between strong feedforward inputs and strong recurrent feedback. However, the theoretical principles governing the efficacy of balanced predictive coding and its robustness to noise, synaptic weight heterogeneity and communication delays remain poorly understood. To discover such principles, we introduce an analytically tractable model of balanced predictive coding, in which the degree of balance and the degree of weight disorder can be dissociated unlike in previous balanced network models, and we develop a mean-field theory of coding accuracy. Overall, our work provides and solves a general theoretical framework for dissecting the differential contributions neural noise, synaptic disorder, chaos, synaptic delays, and balance to the fidelity of predictive neural codes, reveals the fundamental role that balance plays in achieving superclassical scaling, and unifies previously disparate models in theoretical neuroscience.
**********
Interpolation Technique to Speed Up Gradients Propagation in Neural ODEs
Talgat Daulbaev, Alexandr Katrutsa, Larisa Markeeva, Julia Gusak, Andrzej Cichocki, Ivan Oseledets
We propose a simple interpolation-based method for the efficient approximation of gradients in neural ODE models.
We compare it with reverse dynamic method (known in literature as “adjoint method”) to train neural ODEs on classification, density estimation and inference approximation tasks.
We also propose a theoretical justification of our approach using logarithmic norm formalism.
As a result, our method allows faster model training than the reverse dynamic method what was confirmed and validated by extensive numerical experiments for several standard benchmarks.
**********
On the Equivalence between Online and Private Learnability beyond Binary Classification
Young Jung, Baekjin Kim, Ambuj Tewari
Alon et al. [2019] and Bun et al. [2020] recently showed that online learnability and private PAC learnability are equivalent in binary classification. We investigate whether this equivalence extends to multi-class classification and regression. First, we show that private learnability implies online learnability in both settings. Our extension involves studying a novel variant of the Littlestone dimension that depends on a tolerance parameter and on an appropriate generalization of the concept of threshold functions beyond binary classification. Second, we show that while online learnability continues to imply private learnability in multi-class classification, current proof techniques encounter significant hurdles in the regression setting. While the equivalence for regression remains op

en, we provide non-trivial sufficient conditions for an online learnable class t
o also be privately learnable.
**********

AViD Dataset: Anonymized Videos from Diverse Countries
AJ Piergiovanni, Michael Ryoo
We introduce a new public video dataset for action recognition: Anonymized Video
s from Diverse countries (AViD). Unlike existing public video datasets, AViD is
a collection of action videos from many different countries. The motivation is t
o create a public dataset that would benefit training and pretraining of action
recognition models for everybody, rather than making it useful for limited count
ries. Further, all the face identities in the AViD videos are properly anonymize
d to protect their privacy. It also is a static dataset where each video is lice
nsed with the creative commons license. We confirm that most of the existing vid
eo datasets are statistically biased to only capture action videos from a limite
d number of countries. We experimentally illustrate that models trained with suc
h biased datasets do not transfer perfectly to action videos from the other coun
tries, and show that AViD addresses such problem. We also confirm that the new A
ViD dataset could serve as a good dataset for pretraining the models, performing
 comparably or better than prior datasets. The dataset is available at https://g
ithub.com/piergiaj/AViD
**********

Probably Approximately Correct Constrained Learning
Luiz Chamon, Alejandro Ribeiro
As learning solutions reach critical applications in social, industrial, and med
ical domains, the need to curtail their behavior has become paramount. There is
now ample evidence that without explicit tailoring, learning can lead to biased,
 unsafe, and prejudiced solutions. To tackle these problems, we develop a genera
lization theory of constrained learning based on the probably approximately corr
ect (PAC) learning framework. In particular, we show that imposing requirements
does not make a learning problem harder in the sense that any PAC learnable clas
s is also PAC constrained learnable using a constrained counterpart of the empir
ical risk minimization (ERM) rule. For typical parametrized models, however, thi
s learner involves solving a constrained non-convex optimization program for whi
ch even obtaining a feasible solution is challenging. To overcome this issue, we
 prove that under mild conditions the empirical dual problem of constrained lear
ning is also a PAC constrained learner that now leads to a practical constrained
 learning algorithm based solely on solving unconstrained problems. We analyze t
he generalization properties of this solution and use it to illustrate how const
rained learning can address problems in fair and robust classification.
**********

RATT: Recurrent Attention to Transient Tasks for Continual Image Captioning
Riccardo Del Chiaro, Bart■omiej Twardowski, Andrew Bagdanov, Joost van de Weijer
Research on continual learning has led to a variety of approaches to
mitigating catastrophic forgetting in feed-forward classification networks.
Until now surprisingly little attention has been focused on continual learning
of recurrent models applied to problems like image captioning. In this paper
we take a systematic look at continual learning of LSTM-based models for image
captioning. We propose an attention-based approach that explicitly
accommodates the transient nature of vocabularies in continual image
captioning tasks -- i.e. that task vocabularies are not disjoint. We call our
method Recurrent Attention to Transient Tasks (RATT), and also show how to
adapt continual learning approaches based on weight regularization and
knowledge distillation to recurrent continual learning problems. We apply our
approaches to incremental image captioning problem on two new continual
learning benchmarks we define using the MS-COCO and Flickr30 datasets. Our
results demonstrate that RATT is able to sequentially learn five captioning
tasks while incurring no forgetting of previously learned ones.
**********

Decisions, Counterfactual Explanations and Strategic Behavior
Stratis Tsirtsis, Manuel Gomez Rodriguez

As data-driven predictive models are increasingly used to inform decisions, it has been argued that decision makers should provide explanations that help individuals understand what would have to change for these decisions to be beneficial ones. However, there has been little discussion on the possibility that individuals may use the above counterfactual explanations to invest effort strategically and maximize their chances of receiving a beneficial decision. In this paper, our goal is to find policies and counterfactual explanations that are optimal in terms of utility in such a strategic setting. We first show that, given a pre-defined policy, the problem of finding the optimal set of counterfactual explanations is NP-hard. Then, we show that the corresponding objective is nondecreasing and satisfies submodularity and this allows a standard greedy algorithm to enjoy approximation guarantees. In addition, we further show that the problem of jointly finding both the optimal policy and set of counterfactual explanations reduces to maximizing a non-monotone submodular function. As a result, we can use a recent randomized algorithm to solve the problem, which also offers approximation guarantees. Finally, we demonstrate that, by incorporating a matroid constraint into the problem formulation, we can increase the diversity of the optimal set of counterfactual explanations and incentivize individuals across the whole spectrum of the population to self improve. Experiments on synthetic and real lending and credit card data illustrate our theoretical findings and show that the counterfactual explanations and decision policies found by our algorithms achieve higher utility than several competitive baselines.
**********
Hierarchical Patch VAE-GAN: Generating Diverse Videos from a Single Sample
Shir Gur, Sagie Benaim, Lior Wolf
We consider the task of generating diverse and novel videos from a single video sample.
Recently, new hierarchical patch-GAN based approaches were proposed for generating diverse images, given only a single sample at training time. Moving to videos, these approaches fail to generate diverse samples, and often collapse into generating samples similar to the training video. We introduce a novel patch-based variational autoencoder (VAE) which allows for a much greater diversity in generation. Using this tool, a new hierarchical video generation scheme is constructed: at coarse scales, our patch-VAE is employed, ensuring samples are of high diversity. Subsequently, at finer scales, a patch-GAN renders the fine details, resulting in high quality videos.
Our experiments show that the proposed method produces diverse samples in both the image domain, and the more challenging video domain.
Our code and supplementary material (SM) with additional samples are available at https://shirgur.github.io/hp-vae-gan
**********
A Feasible Level Proximal Point Method for Nonconvex Sparse Constrained Optimization
Digvijay Boob, Qi Deng, Guanghui Lan, Yilin Wang
Nonconvex sparse models have received significant attention in high-dimensional machine learning. In this paper, we study a new model consisting of a general convex or nonconvex objectives and a variety of continuous nonconvex sparsity-inducing constraints. For this constrained model, we propose a novel proximal point algorithm that solves a sequence of convex subproblems with gradually relaxed constraint levels. Each subproblem, having a proximal point objective and a convex surrogate constraint, can be efficiently solved based on a fast routine for projection onto the surrogate constraint. We establish the asymptotic convergence of the proposed algorithm to the Karush-Kuhn-Tucker (KKT) solutions. We also establish new convergence complexities to achieve an approximate KKT solution when the objective can be smooth/nonsmooth, deterministic/stochastic and convex/nonconvex with complexity that is on a par with gradient descent for unconstrained optimization problems in respective cases. To the best of our knowledge, this is the first study of the first-order methods with complexity guarantee for nonconvex sparse-constrained problems. We perform numerical experiments to demonstrate the effectiveness of our new model and efficiency of the proposed algorithm for la

rge scale problems.
**********

Reservoir Computing meets Recurrent Kernels and Structured Transforms
Jonathan Dong, Ruben Ohana, Mushegh Rafayelyan, Florent Krzakala
Reservoir Computing is a class of simple yet efficient Recurrent Neural Networks where internal weights are fixed at random and only a linear output layer is trained. In the large size limit, such random neural networks have a deep connection with kernel methods. Our contributions are threefold: a) We rigorously establish the recurrent kernel limit of Reservoir Computing and prove its convergence. b) We test our models on chaotic time series prediction, a classic but challenging benchmark in Reservoir Computing, and show how the Recurrent Kernel is competitive and computationally efficient when the number of data points remains moderate. c) When the number of samples is too large, we leverage the success of structured Random Features for kernel approximation by introducing Structured Reservoir Computing. The two proposed methods, Recurrent Kernel and Structured Reservoir Computing, turn out to be much faster and more memory-efficient than conventional Reservoir Computing.
**********

Comprehensive Attention Self-Distillation for Weakly-Supervised Object Detection
Zeyi Huang, Yang Zou, B. V. K. Vijaya Kumar, Dong Huang
Weakly Supervised Object Detection (WSOD) has emerged as an effective tool to train object detectors using only the image-level category labels. However, without object-level labels, WSOD detectors are prone to detect bounding boxes on salient objects, clustered objects and discriminative object parts. Moreover, the image-level category labels do not enforce consistent object detection across different transformations of the same images. To address the above issues, we propose a Comprehensive Attention Self-Distillation (CASD) training approach for WSOD. To balance feature learning among all object instances, CASD computes the comprehensive attention aggregated from multiple transformations and feature layers of the same images. To enforce consistent spatial supervision on objects, CASD conducts self-distillation on the WSOD networks, such that the comprehensive attention is approximated simultaneously by multiple transformations and feature layers of the same images. CASD produces new state-of-the-art WSOD results on standard benchmarks such as PASCAL VOC 2007/2012 and MS-COCO.
**********

Linear Dynamical Systems as a Core Computational Primitive
Shiva Kaul
Running nonlinear RNNs for T steps takes O(T) time. Our construction, called LDStack, approximately runs them in O(log T) parallel time, and obtains arbitrarily low error via repetition. First, we show nonlinear RNNs can be approximated by a stack of multiple-input, multiple-output (MIMO) LDS. This replaces nonlinearity across time with nonlinearity along depth. Next, we show that MIMO LDS can be approximated by an average or a concatenation of single-input, multiple-output (SIMO) LDS. Finally, we present an algorithm for running (and differentiating) SIMO LDS in O(log T) parallel time. On long sequences, LDStack is much faster than traditional RNNs, yet it achieves similar accuracy in our experiments. Furthermore, LDStack is amenable to linear systems theory. Therefore, it improves not only speed, but also interpretability and mathematical tractability.
**********

Ratio Trace Formulation of Wasserstein Discriminant Analysis
Hexuan Liu, Yunfeng Cai, You-Lin Chen, Ping Li
We reformulate the Wasserstein Discriminant Analysis (WDA) as a ratio trace problem and present an eigensolver-based algorithm to compute the discriminative subspace of WDA. This new formulation, along with the proposed algorithm, can be served as an efficient and more stable alternative to the original trace ratio formulation and its gradient-based algorithm. We provide a rigorous convergence analysis for the proposed algorithm under the self-consistent field framework, which is crucial but missing in the literature. As an application, we combine WDA with low-dimensional clustering techniques, such as K-means, to perform subspace clustering. Numerical experiments on real datasets show promising results of the

ratio trace formulation of WDA in both classification and clustering tasks.
**********

PAC-Bayes Analysis Beyond the Usual Bounds
Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvari, John Shawe-Taylor
We focus on a stochastic learning model where the learner observes a finite set of training examples and the output of the learning process is a data-dependent distribution over a space of hypotheses. The learned data-dependent distribution is then used to make randomized predictions, and the high-level theme addressed here is guaranteeing the quality of predictions on examples that were not seen during training, i.e. generalization. In this setting the unknown quantity of interest is the expected risk of the data-dependent randomized predictor, for which upper bounds can be derived via a PAC-Bayes analysis, leading to PAC-Bayes bounds.
**********

Few-shot Visual Reasoning with Meta-Analogical Contrastive Learning
Youngsung Kim, Jinwoo Shin, Eunho Yang, Sung Ju Hwang
While humans can solve a visual puzzle that requires logical reasoning by observing only few samples, it would require training over a large number of samples for state-of-the-art deep reasoning models to obtain similar performance on the same task. In this work, we propose to solve such a few-shot (or low-shot) abstract visual reasoning problem by resorting to \emph{analogical reasoning}, which is a unique human ability to identify structural or relational similarity between two sets. Specifically, we construct analogical and non-analogical training pairs of two different problem instances, e.g., the latter is created by perturbing or shuffling the original (former) problem. Then, we extract the structural relations among elements in both domains in a pair by enforcing analogical ones to be as similar as possible, while minimizing similarities between non-analogical ones. This analogical contrastive learning allows to effectively learn the relational representations of given abstract reasoning tasks. We validate our method on RAVEN dataset, on which it outperforms state-of-the-art method, with larger gains when the training data is scarce. We further meta-learn our analogical contrastive learning model over the same tasks with diverse attributes, and show that it generalizes to the same visual reasoning problem with unseen attributes.
**********

MPNet: Masked and Permuted Pre-training for Language Understanding
Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu
BERT adopts masked language modeling (MLM) for pre-training and is one of the most successful pre-training models. Since BERT neglects dependency among predicted tokens, XLNet introduces permuted language modeling (PLM) for pre-training to address this problem. However, XLNet does not leverage the full position information of a sentence and thus suffers from position discrepancy between pre-training and fine-tuning. In this paper, we propose MPNet, a novel pre-training method that inherits the advantages of BERT and XLNet and avoids their limitations. MPNet leverages the dependency among predicted tokens through permuted language modeling (vs. MLM in BERT), and takes auxiliary position information as input to make the model see a full sentence and thus reducing the position discrepancy (vs. PLM in XLNet). We pre-train MPNet on a large-scale dataset (over 160GB text corpora) and fine-tune on a variety of down-streaming tasks (GLUE, SQuAD, etc). Experimental results show that MPNet outperforms MLM and PLM by a large margin, and achieves better results on these tasks compared with previous state-of-the-art pre-trained methods (e.g., BERT, XLNet, RoBERTa) under the same model setting. We attach the code in the supplemental materials.
**********

Reinforcement Learning with Feedback Graphs
Christoph Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, Karthik Sridharan
We study RL in the tabular MDP setting where the agent receives additional observations per step in the form of transitions samples. Such additional observations can be provided in many tasks by auxiliary sensors or by leveraging prior knowledge about the environment (e.g., when certain actions yield similar outcome). We formalize this setting using a feedback graph over state-action pairs and sh

ow that model-based algorithms can incorporate additional observations for more sample-efficient learning. We give a regret bound that predominantly depends on the size of the maximum acyclic subgraph of the feedback graph, in contrast with a polynomial dependency on the number of states and actions in the absence of side observations. Finally, we highlight fundamental challenges for leveraging a small dominating set of the feedback graph, as compared to the well-studied bandit setting, and propose a new algorithm that can use such a dominating set to learn a near-optimal policy faster.
**********

Zap Q-Learning With Nonlinear Function Approximation
Shuhang Chen, Adithya M Devraj, Fan Lu, Ana Busic, Sean Meyn
Zap Q-learning is a recent class of reinforcement learning algorithms, motivated primarily as a means to accelerate convergence. Stability theory has been absent outside of two restrictive classes: the tabular setting, and optimal stopping. This paper introduces a new framework for analysis of a more general class of recursive algorithms known as stochastic approximation. Based on this general theory, it is shown that Zap Q-learning is consistent under a non-degeneracy assumption, even when the function approximation architecture is nonlinear. Zap Q-learning with neural network function approximation emerges as a special case, and is tested on examples from OpenAI Gym. Based on multiple experiments with a range of neural network sizes, it is found that the new algorithms converge quickly and are robust to choice of function approximation architecture.
**********

Lipschitz-Certifiable Training with a Tight Outer Bound
Sungyoon Lee, Jaewook Lee, Saerom Park
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Adaptive Non-Monotone Submodular Maximization Subject to a Knapsack Constraint
Georgios Amanatidis, Federico Fusco, Philip Lazos, Stefano Leonardi, Rebecca Reiffenhäuser
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Conformal Symplectic and Relativistic Optimization
Guilherme Franca, Jeremias Sulam, Daniel Robinson, Rene Vidal
Arguably, the two most popular accelerated or momentum-based optimization methods are Nesterov's accelerated gradient and Polyaks's heavy ball, both corresponding to different discretizations of a particular second order differential equation with a friction term. Such connections with continuous-time dynamical systems have been instrumental in demystifying acceleration phenomena in optimization.
Here we study structure-preserving discretizations for a certain class of dissipative (conformal) Hamiltonian systems, allowing us to analyze the symplectic structure of both Nesterov and heavy ball, besides providing several new insights into these methods.
Moreover, we propose a new algorithm based on a dissipative relativistic system that normalizes the momentum and may result in more stable/faster optimization. Importantly, such a method generalizes both Nesterov and heavy ball, each being recovered as distinct limiting cases, and has potential advantages at no additional cost.
**********

Bayes Consistency vs. H-Consistency: The Interplay between Surrogate Loss Functions and the Scoring Function Class
Mingyuan Zhang, Shivani Agarwal

A fundamental question in multiclass classification concerns understanding the c
onsistency properties of surrogate risk minimization algorithms, which minimize
a (often convex) surrogate to the multiclass 0-1 loss. In particular, the framew
ork of calibrated surrogates has played an important role in analyzing the Bayes
 consistency properties of such algorithms, i.e. in studying convergence to a Ba
yes optimal classifier (Zhang, 2004; Tewari and Bartlett, 2007). However, follow
-up work has suggested this framework can be of limited value when studying H-co
nsistency; in particular, concerns have been raised that even when the data come
s from an underlying linear model, minimizing certain convex calibrated surrogat
es over linear scoring functions fails to recover the true model (Long and Serve
dio, 2013). In this paper, we investigate this apparent conundrum. We find that
while some calibrated surrogates can indeed fail to provide H-consistency when m
inimized over a natural-looking but naively chosen scoring function class F, the
 situation can potentially be remedied by minimizing them over a more carefully
chosen class of scoring functions F. In particular, for the popular one-vs-all h
inge and logistic surrogates, both of which are calibrated (and therefore provid
e Bayes consistency) under realizable models, but were previously shown to pose
problems for realizable H-consistency, we derive a form of scoring function clas
s F that enables H-consistency. When H is the class of linear models, the class
F consists of certain piecewise linear scoring functions that are characterized
by the same number of parameters as in the linear case, and minimization over wh
ich can be performed using an adaptation of the min-pooling idea from neural net
work training. Our experiments confirm that the one-vs-all surrogates, when trai
ned over this class of nonlinear scoring functions F, yield better linear multic
lass classifiers than when trained over standard linear scoring functions.
**********

Inverting Gradients - How easy is it to break privacy in federated learning?
Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, Michael Moeller
The idea of federated learning is to collaboratively train a neural network on a
 server. Each user receives the current weights of the network and in turns send
s parameter updates (gradients) based on local data. This protocol has been desi
gned not only to train neural networks data-efficiently, but also to provide pri
vacy benefits for users, as their input data remains on device and only paramete
r gradients are shared.
But how secure is sharing parameter gradients? Previous attacks have provided a
false sense of security, by succeeding only in contrived settings - even for a s
ingle image. However, by exploiting a magnitude-invariant loss along with optimi
zation strategies based on adversarial attacks, we show that is is actually poss
ible to faithfully reconstruct images at high resolution from the knowledge of t
heir parameter gradients, and demonstrate that such a break of privacy is possib
le even for trained deep networks.
We analyze the effects of architecture as well as parameters on the difficulty o
f reconstructing an input image and prove that any input to a fully connected la
yer can be reconstructed analytically independent of the remaining architecture.
 Finally we discuss settings encountered in practice and show that even averagin
g gradients over several iterations or several images does not protect the user'
s privacy in federated learning applications.
**********

Dynamic allocation of limited memory resources in reinforcement learning
Nisheet Patel, Luigi Acerbi, Alexandre Pouget
Biological brains are inherently limited in their capacity to process and store
information, but are nevertheless capable of solving complex tasks with apparent
 ease. Intelligent behavior is related to these limitations, since resource cons
traints drive the need to generalize and assign importance differentially to fea
tures in the environment or memories of past experiences. Recently, there have b
een parallel efforts in reinforcement learning and neuroscience to understand st
rategies adopted by artificial and biological agents to circumvent limitations i
n information storage. However, the two threads have been largely separate. In t
his article, we propose a dynamical framework to maximize expected reward under
constraints of limited resources, which we implement with a cost function that p

enalizes precise representations of action-values in memory, each of which may v
ary in its precision. We derive from first principles an algorithm, Dynamic Reso
urce Allocator (DRA), which we apply to two standard tasks in reinforcement lear
ning and a model-based planning task, and find that it allocates more resources
to items in memory that have a higher impact on cumulative rewards. Moreover, DR
A learns faster when starting with a higher resource budget than what it eventua
lly allocates for performing well on tasks, which may explain why frontal cortic
al areas in biological brains appear more engaged in early stages of learning be
fore settling to lower asymptotic levels of activity. Our work provides a normat
ive solution to the problem of learning how to allocate costly resources to a co
llection of uncertain memories in a manner that is capable of adapting to change
s in the environment.
**********

CryptoNAS: Private Inference on a ReLU Budget
Zahra Ghodsi, Akshaj Kumar Veldanda, Brandon Reagen, Siddharth Garg
Machine learning as a service has given raise to privacy concerns surrounding cl
ients' data and providers' models and has catalyzed research in private inferenc
e (PI): methods to process inferences without disclosing inputs.
Recently, researchers have adapted cryptographic techniques to show PI is possib
le, however all solutions increase inference latency beyond practical limits.
This paper makes the observation that existing models are ill-suited for PI and
proposes a novel NAS method, named CryptoNAS, for finding and tailoring models t
o the needs of PI. The key insight is that in PI operator latency cost are inver
ted:
non-linear operations (e.g., ReLU) dominate latency, while linear layers become
effectively free. We develop the idea of a ReLU budget as a proxy for inference
latency and use CryptoNAS to build models that maximize accuracy within a given
budget. CryptoNAS improves accuracy by 3.4% and  latency by 2.4x over the state-
of-the-art.
**********

A Stochastic Path Integral Differential EstimatoR  Expectation Maximization Algo
rithm
Gersende Fort, Eric Moulines, Hoi-To Wai
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

CHIP: A Hawkes Process Model for Continuous-time Networks with Scalable and Cons
istent Estimation
Makan Arastuie, Subhadeep Paul, Kevin Xu
In many application settings involving networks, such as messages between users
of an on-line social network or transactions between traders in financial market
s, the observed data consist of timestamped relational events, which form a cont
inuous-time network. We propose the Community Hawkes Independent Pairs (CHIP) ge
nerative model for such networks. We show that applying spectral clustering to a
n aggregated adjacency matrix constructed from the CHIP model provides consisten
t community detection for a growing number of nodes and time duration. We also d
evelop consistent and computationally efficient estimators for the model paramet
ers. We demonstrate that our proposed CHIP model and estimation procedure scales
 to large networks with tens of thousands of nodes and provides superior fits th
an existing continuous-time network models on several real networks.
**********

SAC: Accelerating and Structuring Self-Attention via Sparse Adaptive Connection
Xiaoya Li, Yuxian Meng, Mingxin Zhou, Qinghong  Han, Fei Wu, Jiwei Li
While the self-attention mechanism has been widely used in a wide variety of tas
ks, it has the unfortunate property of a quadratic cost with respect to the inpu
t  length, which makes it difficult to deal with long inputs. In this paper, we
present a  method for accelerating and structuring self-attentions: Sparse Adapt
ive Connection (SAC). In SAC, we regard the input sequence as a graph and attent

ion operations are performed between linked nodes. In contrast with previous sel
f-attention models with pre-defined structures (edges), the model learns to cons
truct attention edges to  improve task-specific performances.
In this way, the model is able to select the most salient nodes and reduce the q
uadratic complexity regardless of the sequence length. Based on SAC, we show tha
t previous variants of self-attention models are its special cases. Through exte
nsive experiments on neural machine translation, language modeling, graph repres
entation learning and image classification, we demonstrate SAC is competitive wi
th state-of-the-art models while significantly reducing  memory cost.
**********

Design Space for Graph Neural Networks
Jiaxuan You, Zhitao Ying, Jure Leskovec
The rapid evolution of Graph Neural Networks (GNNs) has led to a growing number
of new architectures as well as novel applications. However, current research fo
cuses on proposing and evaluating specific architectural designs of GNNs, such a
s GCN, GIN, or GAT, as opposed to studying the more general design space of GNNs
 that consists of a Cartesian product of different design dimensions, such as th
e number of layers or the type of the aggregation function. Additionally, GNN de
signs are often specialized to a single task, yet few efforts have been made to
understand how to quickly find the best GNN design for a novel task or a novel d
ataset. Here we define and systematically study the architectural design space f
or GNNs which consists of 315,000 different designs over 32 different predictive
 tasks. Our approach features three key innovations: (1) A general GNN design sp
ace; (2) a GNN task space with a similarity metric, so that for a given novel ta
sk/dataset, we can quickly identify/transfer the best performing architecture; (
3) an efficient and effective design space evaluation method which allows insigh
ts to be distilled from a huge number of model-task combinations. Our key result
s include: (1) A comprehensive set of guidelines for designing well-performing G
NNs; (2) while best GNN designs for different tasks vary significantly, the GNN
task space allows for transferring the best designs across different tasks; (3)
models discovered using our design space achieve state-of-the-art performance. O
verall, our work offers a principled and scalable approach to transition from st
udying individual GNN designs for specific tasks, to systematically studying the
 GNN design space and the task space. Finally, we release GraphGym, a powerful p
latform for exploring different GNN designs and tasks. GraphGym features modular
ized GNN implementation, standardized GNN evaluation, and reproducible and scala
ble experiment management.
**********

HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech
 Synthesis
Jungil Kong, Jaehyeon Kim, Jaekyoung Bae
Several recent work on speech synthesis have employed generative adversarial net
works (GANs) to produce raw waveforms. Although such methods improve the samplin
g efficiency and memory usage, their sample quality has not yet reached that of
autoregressive and flow-based generative models. In this work, we propose HiFi-G
AN, which achieves both efficient and high-fidelity speech synthesis. As speech
audio consists of sinusoidal signals with various periods, we demonstrate that m
odeling periodic patterns of an audio is crucial for enhancing sample quality. A
 subjective human evaluation (mean opinion score, MOS) of a single speaker datas
et indicates that our proposed method demonstrates similarity to human quality w
hile generating 22.05 kHz high-fidelity audio 167.9 times faster than real-time
on a single V100 GPU. We further show the generality of HiFi-GAN to the mel-spec
trogram inversion of unseen speakers and end-to-end speech synthesis. Finally, a
 small footprint version of HiFi-GAN generates samples 13.4 times faster than re
al-time on CPU with comparable quality to an autoregressive counterpart.
**********

 Unbalanced Sobolev Descent
Youssef Mroueh, Mattia Rigotti
We introduce Unbalanced Sobolev Descent (USD), a particle descent algorithm for
transporting a high dimensional source distribution to a target distribution tha

t does not necessarily have the same mass. We define the Sobolev-Fisher discrepancy between distributions and show that it relates to advection-reaction transport equations and the Wasserstein-Fisher-Rao metric between distributions. USD transports particles along gradient flows of the witness function of the Sobolev-Fisher discrepancy (advection step) and reweighs the mass of particles with respect to this witness function (reaction step). The reaction step can be thought of as a birth-death process of the particles with rate of growth proportional to the witness function. When the Sobolev-Fisher witness function is estimated in a Reproducing Kernel Hilbert Space (RKHS), under mild assumptions we show that USD converges asymptotically (in the limit of infinite particles) to the target distribution in the Maximum Mean Discrepancy (MMD) sense. We then give two methods to estimate the Sobolev-Fisher witness with neural networks, resulting in two Neural USD algorithms. The first one implements the reaction step with mirror descent on the weights, while the second implements it through a birth-death process of particles. We show on synthetic examples that USD transports distributions with or without conservation of mass faster than previous particle descent algorithms, and finally demonstrate its use for molecular biology analyses where our method is naturally suited to match developmental stages of populations of differentiating cells based on their single-cell RNA sequencing profile. Code is available at http://github.com/ibm/usd.

**********

Identifying Mislabeled Data using the Area Under the Margin Ranking

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, Kilian Q. Weinberger

Not all data in a typical training set help with generalization; some samples can be overly ambiguous or outrightly mislabeled. This paper introduces a new method to identify such samples and mitigate their impact when training neural networks. At the heart of our algorithm is the Area Under the Margin (AUM) statistic, which exploits differences in the training dynamics of clean and mislabeled samples. A simple procedure - adding an extra class populated with purposefully mislabeled threshold samples - learns a AUM upper bound that isolates mislabeled data. This approach consistently improves upon prior work on synthetic and real-world datasets. On the WebVision50 classification task our method removes 17% of training data, yielding a 1.6% (absolute) improvement in test error. On CIFAR100 removing 13% of the data leads to a 1.2% drop in error.

**********

Combining Deep Reinforcement Learning and Search for Imperfect-Information Games

Noam Brown, Anton Bakhtin, Adam Lerer, Qucheng Gong

The combination of deep reinforcement learning and search at both training and test time is a powerful paradigm that has led to a number of successes in single-agent settings and perfect-information games, best exemplified by AlphaZero. However, prior algorithms of this form cannot cope with imperfect-information games. This paper presents ReBeL, a general framework for self-play reinforcement learning and search that provably converges to a Nash equilibrium in any two-player zero-sum game. In the simpler setting of perfect-information games, ReBeL reduces to an algorithm similar to AlphaZero. Results in two different imperfect-information games show ReBeL converges to an approximate Nash equilibrium. We also show ReBeL achieves superhuman performance in heads-up no-limit Texas hold'em poker, while using far less domain knowledge than any prior poker AI.

**********

High-Throughput Synchronous Deep RL

Iou-Jen Liu, Raymond Yeh, Alexander Schwing

Various parallel actor-learner methods reduce long training times for deep reinforcement learning. Synchronous methods enjoy training stability while having lower data throughput. In contrast, asynchronous methods achieve high throughput but suffer from stability issues and lower sample efficiency due to ‘stale policies.’ To combine the advantages of both methods we propose High-Throughput Synchronous Deep Reinforcement Learning (HTS-RL). In HTS-RL, we perform learning and rollouts concurrently, devise a system design which avoids ‘stale policies’ and ensure that actors interact with environment replicas in an asynchronous manner while maintaining full determinism. We evaluate our approach on Atari games and th

e Google Research Football environment. Compared to synchronous baselines, HTS-R L is 2–6X faster. Compared to state-of-the-art asynchronous methods, HTS-RL has competitive throughput and consistently achieves higher average episode rewards.
**********

Contrastive Learning with Adversarial Examples
Chih-Hui Ho, Nuno Nvasconcelos
Contrastive learning (CL) is a popular technique for self-supervised learning (S SL) of visual representations. It uses pairs of augmentations of unlabeled train ing examples to define a classification task for pretext learning of a deep embe dding. Despite extensive works in augmentation procedures, prior works do not ad dress the selection of challenging negative pairs, as images within a sampled ba tch are treated independently. This paper addresses the problem, by introducing a new family of adversarial examples for constrastive learning and using these e xamples to define a new adversarial training algorithm for SSL, denoted as CLAE. When compared to standard CL, the use of adversarial examples creates more chal lenging positive pairs and adversarial training produces harder negative pairs b y accounting for all images in a batch during the optimization. CLAE is compatib le with many CL methods in the literature. Experiments show that it improves the performance of several existing CL baselines on multiple datasets.
**********

Mixed Hamiltonian Monte Carlo for Mixed Discrete and Continuous Variables
Guangyao Zhou
Hamiltonian Monte Carlo (HMC) has emerged as a powerful Markov Chain Monte Carlo (MCMC) method to sample from complex continuous distributions. However, a funda mental limitation of HMC is that it can not be applied to distributions with mix ed discrete and continuous variables. In this paper, we propose mixed HMC (M-HMC ) as a general framework to address this limitation. M-HMC is a novel family of MCMC algorithms that evolves the discrete and continuous variables in tandem, al lowing more frequent updates of discrete variables while maintaining HMC's abili ty to suppress random-walk behavior. We establish M-HMC's theoretical properties , and present an efficient implementation with Laplace momentum that introduces minimal overhead compared to existing HMC methods. The superior performances of M-HMC over existing methods are demonstrated with numerical experiments on Gauss ian mixture models (GMMs), variable selection in Bayesian logistic regression (B LR), and correlated topic models (CTMs).
**********

Adversarial Sparse Transformer for Time Series Forecasting
Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, Ying Wei, Junzhou Huang
Many approaches have been proposed for time series forecasting, in light of its significance in wide applications including business demand prediction.
However, the existing methods suffer from two key limitations. Firstly, most poi nt prediction models only predict an exact value of each time step without flexi bility, which can hardly capture the stochasticity of data. Even probabilistic p rediction using the likelihood estimation suffers these problems in the same way . Besides, most of them use the auto-regressive generative mode, where ground-tr uth is provided during training and replaced by the network’s own one-step ahead output during inference, causing the error accumulation in inference. Thus they may fail to forecast time series for long time horizon due to the error accumul ation. To solve these issues, in this paper, we propose a new time series foreca sting model -- Adversarial Sparse Transformer (AST), based on Generated Adversar ial Networks (GANs). Specifically, AST adopts a Sparse Transformer as the genera tor to learn a sparse attention map for time series forecasting, and uses a disc riminator to improve the prediction performance from sequence level. Extensive e xperiments on several real-world datasets show the effectiveness and efficiency of our method.
**********

The Surprising Simplicity of the Early-Time Learning Dynamics of Neural Networks
Wei Hu, Lechao Xiao, Ben Adlam, Jeffrey Pennington
Modern neural networks are often regarded as complex black-box functions whose b ehavior is difficult to understand owing to their nonlinear dependence on the da

ta and the nonconvexity in their loss landscapes. In this work, we show that the se common perceptions can be completely false in the early phase of learning. In particular, we formally prove that, for a class of well-behaved input distributions, the early-time learning dynamics of a two-layer fully-connected neural network can be mimicked by training a simple linear model on the inputs. We additionally argue that this surprising simplicity can persist in networks with more layers and with convolutional architecture, which we verify empirically. Key to our analysis is to bound the spectral norm of the difference between the Neural Tangent Kernel (NTK) and an affine transform of the data kernel; however, unlike many previous results utilizing the NTK, we do not require the network to have disproportionately large width, and the network is allowed to escape the kernel regime later in training.

**********

CLEARER: Multi-Scale Neural Architecture Search for Image Restoration

Yuanbiao Gou, Boyun Li, Zitao Liu, Songfan Yang, Xi Peng

Multi-scale neural networks have shown effectiveness in image restoration tasks, which are usually designed and integrated in a handcrafted manner. Different from the existing labor-intensive handcrafted architecture design paradigms, we present a novel method, termed as multi-sCaLe nEural ARchitecture sEarch for image Restoration (CLEARER), which is a speci■cally designed neural architecture search (NAS) for image restoration. Our contributions are twofold. On one hand, we design a multi-scale search space that consists of three task-■exible modules. Namely, 1) Parallel module that connects multi-resolution neural blocks in parallel, while preserving the channels and spatial-resolution in each neural block, 2) Transition module remains the existing multi-resolution features while extending them to a lower resolution, 3) Fusion module integrates multi-resolution features by passing the features of the parallel neural blocks to the current neural blocks. On the other hand, we present novel losses which could 1) balance the tradeoff between the model complexity and performance, which is highly expected to image restoration; and 2) relax the discrete architecture parameters into a continuous distribution which approximates to either 0 or 1. As a result, a differentiable strategy could be employed to search when to fuse or extract multi-resolution features, while the discretization issue faced by the gradient-based NAS could be alleviated. The proposed CLEARER could search a promising architecture in two GPU hours. Extensive experiments show the promising performance of our method comparing with nine image denoising methods and eight image deraining approaches in quantitative and qualitative evaluations. The codes are available at https://github.com/limit-scu.

**********

Hierarchical Gaussian Process Priors for Bayesian Neural Network Weights

Theofanis Karaletsos, Thang D. Bui

Probabilistic neural networks are typically modeled with independent weight priors, which do not capture weight correlations in the prior and do not provide a parsimonious interface to express properties in function space.

A desirable class of priors would represent weights compactly, capture correlations between weights, facilitate calibrated reasoning about uncertainty, and allow inclusion of prior knowledge about the function space such as periodicity or dependence on contexts such as inputs.

To this end, this paper introduces two innovations: (i) a Gaussian process-based hierarchical model for network weights based on unit embeddings that can flexibly encode correlated weight structures, and (ii) input-dependent versions of these weight priors that can provide convenient ways to regularize the function space through the use of kernels defined on contextual inputs.

We show these models provide desirable test-time uncertainty estimates on out-of-distribution data, demonstrate cases of modeling inductive biases for neural networks with kernels which help both interpolation and extrapolation from training data, and demonstrate competitive predictive performance on an active learning benchmark.

**********

Compositional Explanations of Neurons

Jesse Mu, Jacob Andreas
We describe a procedure for explaining neurons in deep representations by identifying compositional logical concepts that closely approximate neuron behavior. Compared to prior work that uses atomic labels as explanations, analyzing neurons compositionally allows us to more precisely and expressively characterize their behavior. We use this procedure to answer several questions on interpretability in models for vision and natural language processing. First, we examine the kinds of abstractions learned by neurons. In image classification, we find that many neurons learn highly abstract but semantically coherent visual concepts, while other polysemantic neurons detect multiple unrelated features; in natural language inference (NLI), neurons learn shallow lexical heuristics from dataset biases. Second, we see whether compositional explanations give us insight into model performance: vision neurons that detect human-interpretable concepts are positively correlated with task performance, while NLI neurons that fire for shallow heuristics are negatively correlated with task performance. Finally, we show how compositional explanations provide an accessible way for end users to produce simple "copy-paste" adversarial examples that change model behavior in predictable ways.
**********

Calibrated Reliable Regression using Maximum Mean Discrepancy
Peng Cui, Wenbo Hu, Jun Zhu
Accurate quantification of uncertainty is crucial for real-world applications of machine learning. However, modern deep neural networks still produce unreliable predictive uncertainty, often yielding over-confident predictions. In this paper, we are concerned with getting well-calibrated predictions in regression tasks. We propose the calibrated regression method using the maximum mean discrepancy by minimizing the kernel embedding measure. Theoretically, the calibration error of our method asymptotically converges to zero when the sample size is large enough. Experiments on non-trivial real datasets show that our method can produce well-calibrated and sharp prediction intervals, which outperforms the related state-of-the-art methods.
**********

Directional convergence and alignment in deep learning
Ziwei Ji, Matus Telgarsky
In this paper, we show that although the minimizers of cross-entropy and related classification losses are off at infinity, network weights learned by gradient flow converge in direction, with an immediate corollary that network predictions, training errors, and the margin distribution also converge. This proof holds for deep homogeneous networks – a broad class of networks allowing for ReLU, max-pooling, linear, and convolutional layers – and we additionally provide empirical support not just close to the theory (e.g., the AlexNet), but also on non-homogeneous networks (e.g., the DenseNet). If the network further has locally Lipschitz gradients, we show that these gradients also converge in direction, and asymptotically align with the gradient flow path, with consequences on margin maximization, convergence of saliency maps, and a few other settings. Our analysis complements and is distinct from the well-known neural tangent and mean-field theories, and in particular makes no requirements on network width and initialization, instead merely requiring perfect classification accuracy. The proof proceeds by developing a theory of unbounded nonsmooth Kurdyka-■ojasiewicz inequalities for functions definable in an o-minimal structure, and is also applicable outside deep learning.
**********

Functional Regularization for Representation Learning: A Unified Theoretical Perspective
Siddhant Garg, Yingyu Liang
Unsupervised and self-supervised learning approaches have become a crucial tool to learn representations for downstream prediction tasks. While these approaches are widely used in practice and achieve impressive empirical gains, their theoretical understanding largely lags behind. Towards bridging this gap, we present a unifying perspective where several such approaches can be viewed as imposing a

regularization on the representation via a learnable function using unlabeled data. We propose a discriminative theoretical framework for analyzing the sample complexity of these approaches, which generalizes the framework of (Balcan and Blum, 2010) to allow learnable regularization functions. Our sample complexity bounds show that, with carefully chosen hypothesis classes to exploit the structure in the data, these learnable regularization functions can prune the hypothesis space, and help reduce the amount of labeled data needed. We then provide two concrete examples of functional regularization, one using auto-encoders and the other using masked self-supervision, and apply our framework to quantify the reduction in the sample complexity bound of labeled data. We also provide complementary empirical results to support our analysis.

**********

Provably Efficient Online Hyperparameter Optimization with Population-Based Bandits

Jack Parker-Holder, Vu Nguyen, Stephen J. Roberts

Many of the recent triumphs in machine learning are dependent on well-tuned hyperparameters. This is particularly prominent in reinforcement learning (RL) where a small change in the configuration can lead to failure. Despite the importance of tuning hyperparameters, it remains expensive and is often done in a naive and laborious way. A recent solution to this problem is Population Based Training (PBT) which updates both weights and hyperparameters in a \emph{single training run} of a population of agents. PBT has been shown to be particularly effective in RL, leading to widespread use in the field. However, PBT lacks theoretical guarantees since it relies on random heuristics to explore the hyperparameter space. This inefficiency means it typically requires vast computational resources, which is prohibitive for many small and medium sized labs. In this work, we introduce the first provably efficient PBT-style algorithm, Population-Based Bandits (PB2). PB2 uses a probabilistic model to guide the search in an efficient way, making it possible to discover high performing hyperparameter configurations with far fewer agents than typically required by PBT. We show in a series of RL experiments that PB2 is able to achieve high performance with a modest computational budget.

**********

Understanding Global Feature Contributions With Additive Importance Measures

Ian Covert, Scott M. Lundberg, Su-In Lee

Understanding the inner workings of complex machine learning models is a long-standing problem and most recent research has focused on local interpretability. To assess the role of individual input features in a global sense, we explore the perspective of defining feature importance through the predictive power associated with each feature. We introduce two notions of predictive power (model-based and universal) and formalize this approach with a framework of additive importance measures, which unifies numerous methods in the literature. We then propose SAGE, a model-agnostic method that quantifies predictive power while accounting for feature interactions. Our experiments show that SAGE can be calculated efficiently and that it assigns more accurate importance values than other methods.

**********

Online Non-Convex Optimization with Imperfect Feedback

Amélie Héliou, Matthieu Martin, Panayotis Mertikopoulos, Thibaud Rahier

We consider the problem of online learning with non-convex losses. In terms of feedback, we assume that the learner observes – or otherwise constructs – an inexact model for the loss function encountered at each stage, and we propose a mixed-strategy learning policy based on dual averaging. In this general context, we derive a series of tight regret minimization guarantees, both for the learner’s static (external) regret, as well as the regret incurred against the best dynamic policy in hindsight. Subsequently, we apply this general template to the case where the learner only has access to the actual loss incurred at each stage of the process. This is achieved by means of a kernel-based estimator which generates an inexact model for each round’s loss function using only the learner’s realized losses as input.

**********

Co-Tuning for Transfer Learning
Kaichao You, Zhi Kou, Mingsheng Long, Jianmin 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.
**********

Multifaceted Uncertainty Estimation for Label-Efficient Deep Learning
Weishi Shi, Xujiang Zhao, Feng Chen, Qi Yu
We present a novel multi-source uncertainty prediction approach that enables deep learning (DL) models to be actively trained with much less labeled data. By leveraging the second-order uncertainty representation provided by subjective logic (SL), we conduct evidence-based theoretical analysis and formally decompose the predicted entropy over multiple classes into two distinct sources of uncertainty: vacuity and dissonance, caused by lack of evidence and conflict of strong evidence, respectively. The evidence based entropy decomposition provides deeper insights on the nature of uncertainty, which can help effectively explore a large and high-dimensional unlabeled data space. We develop a novel loss function that augments DL based evidence prediction with uncertainty anchor sample identification. The accurately estimated multiple sources of uncertainty are systematically integrated and dynamically balanced using a data sampling function for label-efficient active deep learning (ADL). Experiments conducted over both synthetic and real data and comparison with competitive AL methods demonstrate the effectiveness of the proposed ADL model.
**********

Continuous Surface Embeddings
Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, Andrea Vedaldi
In this work, we focus on the task of learning and representing dense correspondences in deformable object categories. While this problem has been considered before, solutions so far have been rather ad-hoc for specific object types (i.e., humans), often with significant manual work involved. However, scaling the geometry understanding to all objects in nature requires more automated approaches that can also express correspondences between related, but geometrically different objects. To this end, we propose a new, learnable image-based representation of dense correspondences. Our model predicts, for each pixel in a 2D image, an embedding vector of the corresponding vertex in the object mesh, therefore establishing dense correspondences between image pixels and 3D object geometry. We demonstrate that the proposed approach performs on par or better than the state-of-the-art methods for dense pose estimation for humans, while being conceptually simpler. We also collect a new in-the-wild dataset of dense correspondences for animal classes and demonstrate that our framework scales naturally to the new deformable object categories.
**********

Succinct and Robust Multi-Agent Communication With Temporal Message Control
Sai Qian Zhang, Qi Zhang, Jieyu Lin
Recent studies have shown that introducing communication between agents can significantly improve overall performance in cooperative Multi-agent reinforcement learning (MARL). However, existing communication schemes often require agents to exchange an excessive number of messages at run-time under a reliable communication channel, which hinders its practicality in many real-world situations. In this paper, we present \textit{Temporal Message Control} (TMC), a simple yet effective approach for achieving succinct and robust communication in MARL. TMC applies a temporal smoothing technique to drastically reduce the amount of information exchanged between agents. Experiments show that TMC can significantly reduce inter-agent communication overhead without impacting accuracy. Furthermore, TMC demonstrates much better robustness against transmission loss than existing approaches in lossy networking environments.
**********

Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 Execution Engines: Learning to Execute Subroutines
Yujun Yan, Kevin Swersky, Danai Koutra, Parthasarathy Ranganathan, Milad Hashemi
A significant effort has been made to train neural networks that replicate algorithmic reasoning, but they often fail to learn the abstract concepts underlying these algorithms. This is evidenced by their inability to generalize to data distributions that are outside of their restricted training sets, namely larger inputs and unseen data. We study these generalization issues at the level of numerical subroutines that comprise common algorithms like sorting, shortest paths, and minimum spanning trees. First, we observe that transformer-based sequence-to-sequence models can learn subroutines like sorting a list of numbers, but their performance rapidly degrades as the length of lists grows beyond those found in the training set. We demonstrate that this is due to attention weights that lose fidelity with longer sequences, particularly when the input numbers are numerically similar. To address the issue, we propose a learned conditional masking mechanism, which enables the model to strongly generalize far outside of its training range with near-perfect accuracy on a variety of algorithms. Second, to generalize to unseen data, we show that encoding numbers with a binary representation leads to embeddings with rich structure once trained on downstream tasks like addition or multiplication. This allows the embedding to handle missing data by faithfully interpolating numbers not seen during training.
**********

Random Reshuffling: Simple Analysis with Vast Improvements
Konstantin Mishchenko, Ahmed Khaled, Peter Richtarik
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Long-Horizon Visual Planning with Goal-Conditioned Hierarchical Predictors
Karl Pertsch, Oleh Rybkin, Frederik Ebert, Shenghao Zhou, Dinesh Jayaraman, Chelsea Finn, Sergey Levine
The ability to predict and plan into the future is fundamental for agents acting in the world. To reach a faraway goal, we predict trajectories at multiple time scales, first devising a coarse plan towards the goal and then gradually filling in details. In contrast, current learning approaches for visual prediction and planning fail on long-horizon tasks as they generate predictions (1)~without considering goal information, and (2)~at the finest temporal resolution, one step at a time. In this work we propose a framework for visual prediction and planning that is able to overcome both of these limitations. First, we formulate the problem of predicting towards a goal and propose the corresponding class of latent space goal-conditioned predictors (GCPs). GCPs significantly improve planning efficiency by constraining the search space to only those trajectories that reach the goal. Further, we show how GCPs can be naturally formulated as hierarchical models that, given two observations, predict an observation between them, and by recursively subdividing each part of the trajectory generate complete sequences. This divide-and-conquer strategy is effective at long-term prediction, and enables us to design an effective hierarchical planning algorithm that optimizes trajectories in a coarse-to-fine manner. We show that by using both goal-conditioning and hierarchical prediction, GCPs enable us to solve visual planning tasks with much longer horizon than previously possible. See prediction and planning videos on the supplementary website: sites.google.com/view/video-gcp.
**********

Statistical Optimal Transport posed as Learning Kernel Embedding
Saketha Nath Jagarlapudi, Pratik Kumar Jawanpuria
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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-Resolution Correspondence Networks
Xinghui Li, Kai Han, Shuda Li, Victor Prisacariu
We tackle the problem of establishing dense pixel-wise correspondences between a pair of images. In this work, we introduce Dual-Resolution Correspondence Networks (DualRC-Net), to obtain pixel-wise correspondences in a coarse-to-fine manner. DualRC-Net extracts both coarse- and fine- resolution feature maps. The coarse maps are used to produce a full but coarse 4D correlation tensor, which is then refined by a learnable neighbourhood consensus module. The fine-resolution feature maps are used to obtain the final dense correspondences guided by the refined coarse 4D correlation tensor. The selected coarse-resolution matching scores allow the fine-resolution features to focus only on a limited number of possible matches with high confidence. In this way, DualRC-Net dramatically increases matching reliability and localisation accuracy, while avoiding to apply the expensive 4D convolution kernels on fine-resolution feature maps. We comprehensively evaluate our method on large-scale public benchmarks including HPatches, InLoc, and Aachen Day-Night. It achieves state-of-the-art results on all of them.
**********

Advances in Black-Box VI: Normalizing Flows, Importance Weighting, and Optimization
Abhinav Agrawal, Daniel R. Sheldon, Justin Domke
Recent research has seen several advances relevant to black-box VI, but the current state of automatic posterior inference is unclear. One such advance is the use of normalizing flows to define flexible posterior densities for deep latent variable models. Another direction is the integration of Monte-Carlo methods to serve two purposes; first, to obtain tighter variational objectives for optimization, and second, to define enriched variational families through sampling. However, both flows and variational Monte-Carlo methods remain relatively unexplored for black-box VI. Moreover, on a pragmatic front, there are several optimization considerations like step-size scheme, parameter initialization, and choice of gradient estimators, for which there are no clear guidance in the existing literature. In this paper, we postulate that black-box VI is best addressed through a careful combination of numerous algorithmic components. We evaluate components relating to optimization, flows, and Monte-Carlo methods on a benchmark of 30 models from the Stan model library. The combination of these algorithmic components significantly advances the state-of-the-art "out of the box" variational inference.
**********

f-Divergence Variational Inference
Neng Wan, Dapeng Li, NAIRA HOVAKIMYAN
This paper introduces the f-divergence variational inference (f-VI) that generalizes variational inference to all f-divergences. Initiated from minimizing a crafty surrogate f-divergence that shares the statistical consistency with the f-divergence, the f-VI framework not only unifies a number of existing VI methods, e.g. Kullback–Leibler VI, Renyi's alpha-VI, and chi-VI, but offers a standardized toolkit for VI subject to arbitrary divergences from f-divergence family. A general f-variational bound is derived and provides a sandwich estimate of marginal likelihood (or evidence). The development of the f-VI unfolds with a stochastic optimization scheme that utilizes the reparameterization trick, importance weighting and Monte Carlo approximation; a mean-field approximation scheme that generalizes the well-known coordinate ascent variational inference (CAVI) is also proposed for f-VI. Empirical examples, including variational autoencoders and Bayesian neural networks, are provided to demonstrate the effectiveness and the wide applicability of f-VI.

**********
Unfolding recurrence by Green’s functions for optimized reservoir computing
Sandra Nestler, Christian Keup, David Dahmen, Matthieu Gilson, Holger Rauhut, Moritz Helias
Cortical networks are strongly recurrent, and neurons have intrinsic temporal dynamics. This sets them apart from deep feed-forward networks. Despite the tremendous progress in the application of deep feed-forward networks and their theoretical understanding, it remains unclear how the interplay of recurrence and non-linearities in recurrent cortical networks contributes to their function. The purpose of this work is to present a solvable recurrent network model that links to feed forward networks. By perturbative methods we transform the time-continuous, recurrent dynamics into an effective feed-forward structure of linear and non-linear temporal kernels. The resulting analytical expressions allow us to build optimal time-series classifiers from random reservoir networks. Firstly, this allows us to optimize not only the readout vectors, but also the input projection, demonstrating a strong potential performance gain. Secondly, the analysis exposes how the second order stimulus statistics is a crucial element that interacts with the non-linearity of the dynamics and boosts performance.
**********
The Dilemma of TriHard Loss and an Element-Weighted TriHard Loss for Person Re-Identification
Yihao Lv, Youzhi Gu, Liu Xinggao
Triplet loss with batch hard mining (TriHard loss) is an important variation of triplet loss inspired by the idea that hard triplets improve the performance of metric leaning networks. However, there is a dilemma in the training process. The hard negative samples contain various quite similar characteristics compared with anchors and positive samples in a batch. Features of these characteristics should be clustered between anchors and positive samples while are also utilized to repel between anchors and hard negative samples. It is harmful for learning mutual features within classes. Several methods to alleviate the dilemma are designed and tested. In the meanwhile, an element-weighted TriHard loss is emphatically proposed to enlarge the distance between partial elements of feature vectors selectively which represent the different characteristics between anchors and hard negative samples. Extensive evaluations are conducted on Market1501 and MSMT17 datasets and the results achieve state-of-the-art on public baselines.
**********
Disentangling by Subspace Diffusion
David Pfau, Irina Higgins, Alex Botev, Sébastien Racanière
We present a novel nonparametric algorithm for symmetry-based disentangling of data manifolds, the Geometric Manifold Component Estimator (GEOMANCER). GEOMANCER provides a partial answer to the question posed by Higgins et al.(2018): is it possible to learn how to factorize a Lie group solely from observations of the orbit of an object it acts on? We show that fully unsupervised factorization of a data manifold is possible if the true metric of the manifold is known and each factor manifold has nontrivial holonomy – for example, rotation in 3D. Our algorithm works by estimating the subspaces that are invariant under random walk diffusion, giving an approximation to the de Rham decomposition from differential geometry. We demonstrate the efficacy of GEOMANCER on several complex synthetic manifolds. Our work reduces the question of whether unsupervised disentangling is possible to the question of whether unsupervised metric learning is possible, providing a unifying insight into the geometric nature of representation learning.
**********
Towards Neural Programming Interfaces
Zachary Brown, Nathaniel Robinson, David Wingate, Nancy Fulda
It is notoriously difficult to control the behavior of artificial neural networks such as generative neural language models. We recast the problem of controlling natural language generation as that of learning to interface with a pretrained language model, just as Application Programming Interfaces (APIs) control the behavior of programs by altering hyperparameters. In this new paradigm, a special

ized neural network (called a Neural Programming Interface or NPI) learns to int
erface with a pretrained language model by manipulating the hidden activations o
f the pretrained model to produce desired outputs.  Importantly, no permanent ch
anges are made to the weights of the original model, allowing us to re-purpose p
retrained models for new tasks without overwriting any aspect of the language mo
del. We also contribute a new data set construction algorithm and GAN-inspired l
oss function that allows us to train NPI models to control outputs of autoregres
sive transformers. In experiments against other state-of-the-art approaches, we
demonstrate the efficacy of our methods using OpenAI’s GPT-2 model, successfully
 controlling noun selection, topic aversion, offensive speech filtering, and oth
er aspects of language while largely maintaining the controlled model's fluency
under deterministic settings.
**********

Discovering Symbolic Models from Deep Learning with Inductive Biases
Miles Cranmer, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, D
avid Spergel, Shirley Ho
We develop a general approach to distill symbolic representations of a learned d
eep model by introducing strong inductive biases. We focus on Graph Neural Netwo
rks (GNNs). The technique works as follows: we first encourage sparse latent rep
resentations when we train a GNN in a supervised setting, then we apply symbolic
 regression to components of the learned model to extract explicit physical rela
tions. We find the correct known equations, including force laws and Hamiltonian
s, can be extracted from the neural network. We then apply our method to a non-t
rivial cosmology example—a detailed dark matter simulation—and discover a new an
alytic formula which can predict the concentration of dark matter from the mass
distribution of nearby cosmic structures. The symbolic expressions extracted fro
m the GNN using our technique also generalized to out-of-distribution-data bette
r than the GNN itself. Our approach offers alternative directions for interpreti
ng neural networks and discovering novel physical principles from the representa
tions they learn.
**********

Real World Games Look Like Spinning Tops
Wojciech M. Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omid
shafiei, David Balduzzi, Max Jaderberg
This paper investigates the geometrical properties of real world games (e.g. Tic
-Tac-Toe, Go, StarCraft II).
We hypothesise that their geometrical structure resembles a spinning top, with t
he upright axis representing transitive strength, and the radial axis representi
ng the non-transitive dimension, which corresponds to the number of cycles that
exist at a particular transitive strength.
We prove the existence of this geometry for a wide class of real world games by
exposing their temporal nature.
Additionally, we show that this unique structure also has consequences for learn
ing - it clarifies why populations of strategies are necessary for training of a
gents, and how population size relates to the structure of the game.
Finally, we empirically validate these claims by using a selection of nine real
world two-player zero-sum symmetric games, showing 1) the spinning top structure
 is revealed and can be easily reconstructed by using a new method of Nash clust
ering to measure the interaction between transitive and cyclical strategy behavi
our, and 2) the effect that population size has on the convergence of learning i
n these games.
**********

Cooperative Heterogeneous Deep Reinforcement Learning
Han Zheng, Pengfei Wei, Jing Jiang, Guodong Long, Qinghua Lu, Chengqi Zhang
Numerous deep reinforcement learning agents have been proposed, and each
of them has its strengths and flaws. In this work, we present a Cooperative
Heterogeneous Deep Reinforcement Learning (CHDRL) framework that can learn
a policy by integrating the advantages of heterogeneous agents. Specifically, we
propose a cooperative learning framework that classifies heterogeneous agents in
to two classes: global agents and local agents. Global agents are off-policy age

nts that can utilize experiences from the other agents. Local agents are either on-policy agents or population-based evolutionary algorithms (EAs) agents that can explore the local area effectively. We employ global agents, which are sample-efficient, to guide the learning of local agents so that local agents can benefit from the sample-efficient agents and simultaneously maintain their advantages, e.g., stability. Global agents also benefit from effective local searches. Experimental studies on a range of continuous control tasks from the Mujoco benchmark show that CHDRL achieves better performance compared with state-of-the-art baselines.
**********

Mitigating Forgetting in Online Continual Learning via Instance-Aware Parameterization

Hung-Jen Chen, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, Min Sun

Online continual learning is a challenging scenario where a model needs to learn from a continuous stream of data without revisiting any previously encountered data instances. The phenomenon of catastrophic forgetting is worsened since the model should not only address the forgetting at the task-level but also at the data instance-level within the same task. To mitigate this, we leverage the concept of "instance awareness" in the neural network, where each data instance is classified by a path in the network searched by the controller from a meta-graph. To preserve the knowledge we learn from previous instances, we proposed a method to protect the path by restricting the gradient updates of one instance from overriding past updates calculated from previous instances if these instances are not similar. On the other hand, it also encourages fine-tuning the path if the incoming instance shares the similarity with previous instances. The mechanism of selecting paths according to instances similarity is naturally determined by the controller, which is compact and online updated. Experimental results show that the proposed method outperforms state-of-the-arts in online continual learning. Furthermore, the proposed method is evaluated against a realistic setting where the boundaries between tasks are blurred. Experimental results confirm that the proposed method outperforms the state-of-the-arts on CIFAR-10, CIFAR-100, and Tiny-ImageNet.
**********

ImpatientCapsAndRuns: Approximately Optimal Algorithm Configuration from an Infinite Pool

Gellert Weisz, András György, Wei-I Lin, Devon Graham, Kevin Leyton-Brown, Csaba Szepesvari, Brendan Lucier

Algorithm configuration procedures optimize parameters of a given algorithm to perform well over a distribution of inputs. Recent theoretical work focused on the case of selecting between a small number of alternatives. In practice, parameter spaces are often very large or infinite, and so successful heuristic procedures discard parameters ``impatiently'', based on very few observations. Inspired by this idea, we introduce ImpatientCapsAndRuns, which quickly discards less promising configurations, significantly speeding up the search procedure compared to previous algorithms with theoretical guarantees, while still achieving optimal runtime up to logarithmic factors under mild assumptions. Experimental results demonstrate a practical improvement.
**********

Dense Correspondences between Human Bodies via Learning Transformation Synchronization on Graphs

Xiangru Huang, Haitao Yang, Etienne Vouga, Qixing Huang

We introduce an approach for establishing dense correspondences between partial scans of human models and a complete template model. Our approach's key novelty lies in formulating dense correspondence computation as initializing and synchronizing local transformations between the scan and the template model. We introduce an optimization formulation for synchronizing transformations among a graph of the input scan, which automatically enforces smoothness of correspondences and recovers the underlying articulated deformations. We then show how to convert the iterative optimization procedure among a graph of the input scan into an end-to-end trainable network. The network design utilizes additional trainable param

eters to break the barrier of the original optimization formulation's exact and robust recovery conditions. Experimental results on benchmark datasets demonstrate that our approach considerably outperforms baseline approaches in accuracy and robustness.
**********
Reasoning about Uncertainties in Discrete-Time Dynamical Systems using Polynomial Forms.
Sriram Sankaranarayanan, Yi Chou, Eric Goubault, Sylvie Putot
In this paper, we propose polynomial forms to represent distributions of state variables over time for discrete-time stochastic dynamical systems. This problem arises in a variety of applications in areas ranging from biology to robotics. Our approach allows us to rigorously represent the probability distribution of state variables over time, and provide guaranteed bounds on the expectations, moments and probabilities of tail events involving the state variables. First, we recall ideas from interval arithmetic, and use them to rigorously represent the state variables at time t as a function of the initial state variables and noise symbols that model the random exogenous inputs encountered before time t. Next, we show how concentration of measure inequalities can be employed to prove rigorous bounds on the tail probabilities of these state variables. We demonstrate interesting applications that demonstrate how our approach can be useful in some situations to establish mathematically guaranteed bounds that are of a different nature from those obtained through simulations with pseudo-random numbers.
**********
Applications of Common Entropy for Causal Inference
Murat Kocaoglu, Sanjay Shakkottai, Alexandros G. Dimakis, Constantine Caramanis, Sriram Vishwanath
We study the problem of discovering the simplest latent variable that can make two observed discrete variables conditionally independent. The minimum entropy required for such a latent is known as common entropy in information theory. We extend this notion to Renyi common entropy by minimizing the Renyi entropy of the latent variable. To efficiently compute common entropy, we propose an iterative algorithm that can be used to discover the trade-off between the entropy of the latent variable and the conditional mutual information of the observed variables. We show two applications of common entropy in causal inference: First, under the assumption that there are no low-entropy mediators, it can be used to distinguish direct causation from spurious correlation among almost all joint distributions on simple causal graphs with two observed variables. Second, common entropy can be used to improve constraint-based methods such as PC or FCI algorithms in the small-sample regime, where these methods are known to struggle. We propose a modification to these constraint-based methods to assess if a separating set found by these algorithms are valid using common entropy. We finally evaluate our algorithms on synthetic and real data to establish their performance.
**********
SGD with shuffling: optimal rates without component convexity and large epoch requirements
Kwangjun Ahn, Chulhee Yun, Suvrit Sra
We study without-replacement SGD for solving finite-sum optimization problems. Specifically, depending on how the indices of the finite-sum are shuffled, we consider the RandomShuffle (shuffle at the beginning of each epoch) and SingleShuffle (shuffle only once) algorithms. First, we establish minimax optimal convergence rates of these algorithms up to poly-log factors. Notably, our analysis is general enough to cover gradient dominated nonconvex costs, and does not rely on the convexity of individual component functions unlike existing optimal convergence results. Secondly, assuming convexity of the individual components, we further sharpen the tight convergence results for RandomShuffle by removing the drawbacks common to all prior arts: large number of epochs required for the results to hold, and extra poly-log factor gaps to the lower bound.
**********

Unsupervised Joint k-node Graph Representations with Compositional Energy-Based Models
Leonardo Cotta, Carlos H. C. Teixeira, Ananthram Swami, Bruno Ribeiro
Existing Graph Neural Network (GNN) methods that learn inductive unsupervised graph representations focus on learning node and edge representations by predicting observed edges in the graph. Although such approaches have shown advances in downstream node classification tasks, they are ineffective in jointly representing larger k-node sets, k{>}2. We propose MHM-GNN, an inductive unsupervised graph representation approach that combines joint k-node representations with energy-based models (hypergraph Markov networks) and GNNs. To address the intractability of the loss that arises from this combination, we endow our optimization with a loss upper bound using a finite-sample unbiased Markov Chain Monte Carlo estimator. Our experiments show that the unsupervised joint k-node representations of MHM-GNN produce better unsupervised representations than existing approaches from the literature.
**********
Neural Manifold Ordinary Differential Equations
Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, Christopher M. De Sa
To better conform to data geometry, recent deep generative modelling techniques adapt Euclidean constructions to non-Euclidean spaces. In this paper, we study normalizing flows on manifolds. Previous work has developed flow models for specific cases; however, these advancements hand craft layers on a manifold-by-manifold basis, restricting generality and inducing cumbersome design constraints. We overcome these issues by introducing Neural Manifold Ordinary Differential Equations, a manifold generalization of Neural ODEs, which enables the construction of Manifold Continuous Normalizing Flows (MCNFs). MCNFs require only local geometry (therefore generalizing to arbitrary manifolds) and compute probabilities with continuous change of variables (allowing for a simple and expressive flow construction). We find that leveraging continuous manifold dynamics produces a marked improvement for both density estimation and downstream tasks.
**********
CO-Optimal Transport
Vayer Titouan, Ievgen Redko, Rémi Flamary, Nicolas Courty
Optimal transport (OT) is a powerful geometric and probabilistic tool for finding correspondences and measuring similarity between two distributions. Yet, its original formulation relies on the existence of a cost function between the samples of the two distributions, which makes it impractical when they are supported on different spaces. To circumvent this limitation, we propose a novel OT problem, named COOT for CO-Optimal Transport, that simultaneously optimizes two transport maps between both samples and features, contrary to other approaches that either discard the individual features by focusing on pairwise distances between samples or need to model explicitly the relations between them. We provide a thorough theoretical analysis of our problem, establish its rich connections with other OT-based distances and demonstrate its versatility with two machine learning applications in heterogeneous domain adaptation and co-clustering/data summarization, where COOT leads to performance improvements over the state-of-the-art methods.
**********
Continuous Meta-Learning without Tasks
James Harrison, Apoorva Sharma, Chelsea Finn, Marco Pavone
Meta-learning is a promising strategy for learning to efficiently learn using data gathered from a distribution of tasks. However, the meta-learning literature thus far has focused on the task segmented setting, where at train-time, offline data is assumed to be split according to the underlying task, and at test-time, the algorithms are optimized to learn in a single task. In this work, we enable the application of generic meta-learning algorithms to settings where this task segmentation is unavailable, such as continual online learning with unsegmented time series data. We present meta-learning via online changepoint analysis (MOCA), an approach which augments a meta-learning algorithm with a differentiable

Bayesian changepoint detection scheme. The framework allows both training and te
sting directly on time series data without segmenting it into discrete tasks. We
 demonstrate the utility of this approach on three nonlinear meta-regression ben
chmarks as well as two meta-image-classification benchmarks.
**********

A mathematical theory of cooperative communication
Pei Wang, Junqi Wang, Pushpi Paranamana, Patrick Shafto
Cooperative communication plays a central role in theories of human cognition, l
anguage, development, culture, and human-robot interaction. Prior models of coop
erative communication are algorithmic in nature and do not shed light on why coo
peration may yield effective belief transmission and what limitations may arise
due to differences between beliefs of agents. Through a connection to the theory
 of optimal transport, we establishing a mathematical framework for cooperative
communication. We derive prior models as special cases, statistical interpretati
ons of belief transfer plans, and proofs of robustness and instability. Computat
ional simulations support and elaborate our theoretical results, and demonstrate
 fit to human behavior. The results show that cooperative communication provably
 enables effective, robust belief transmission which is required to explain feat
s of human learning and improve human-machine interaction.
**********

Penalized Langevin dynamics with vanishing penalty  for smooth and log-concave t
argets
Avetik Karagulyan, Arnak Dalalyan
Requests for name changes in the electronic proceedings will be accepted with 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 Invariances in Neural Networks from Training Data
Gregory Benton, Marc Finzi, Pavel Izmailov, Andrew G. Wilson
Invariances to translations have imbued convolutional neural networks with power
ful generalization properties. However, we often do not know a priori what invar
iances are present in the data, or to what extent a model should be invariant to
 a given augmentation. We show how to learn invariances by parameterizing a dist
ribution over augmentations and optimizing the training loss simultaneously with
 respect to the network parameters and augmentation parameters. With this simple
 procedure we can recover the correct set and extent of invariances on image cla
ssification, regression, segmentation, and molecular property prediction from a
large space of augmentations, on training data alone. We show our approach is co
mpetitive with methods that are specialized to each task with the appropriate ha
rd-coded invariances, without providing any prior knowledge of which invariance
is needed.
**********

A Finite-Time Analysis of Two Time-Scale Actor-Critic Methods
Yue Frank Wu, Weitong ZHANG, Pan Xu, Quanquan Gu
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Pruning Filter in Filter
Fanxu Meng, Hao Cheng, Ke Li, Huixiang Luo, Xiaowei Guo, Guangming Lu, Xing Sun
Pruning has become a very powerful and effective technique to compress and accel
erate modern neural networks. Existing pruning methods can be grouped into two c
ategories: filter pruning (FP) and weight pruning (WP). FP wins at hardware comp
atibility but loses at the compression ratio compared with WP. To converge the s
trength of both methods, we propose to prune the filter in the filter. Specifica
lly, we treat a filter F, whose size is CKK, as KK stripes, i.e., 11 filters, th
en by pruning the stripes instead of the whole filter, we can achieves finer gra
nularity than traditional FP while being hardware friendly. We term our method a

s SWP (Stripe-Wise Pruning). SWP is implemented by introducing a novel learnable matrix called Filter Skeleton, whose values reflect the optimal shape of each filter. As some recent work has shown that the pruned architecture is more crucial than the inherited important weights, we argue that the architecture of a single filter, i.e., the Filter Skeleton, also matters. Through extensive experiments, we demonstrate that SWP is more effective compared to the previous FP-based methods and achieves the state-of-art pruning ratio on CIFAR-10 and ImageNet datasets without obvious accuracy drop.

**********

Learning to Mutate with Hypergradient Guided Population

Zhiqiang Tao, Yaliang Li, Bolin Ding, Ce Zhang, Jingren Zhou, Yun Fu

Computing the gradient of model hyperparameters, i.e., hypergradient, enables a promising and natural way to solve the hyperparameter optimization task. However, gradient-based methods could lead to suboptimal solutions due to the non-convex nature of optimization in a complex hyperparameter space. In this study, we propose a hyperparameter mutation (HPM) algorithm to explicitly consider a learnable trade-off between using global and local search, where we adopt a population of student models to simultaneously explore the hyperparameter space guided by hypergradient and leverage a teacher model to mutate the underperforming students by exploiting the top ones. The teacher model is implemented with an attention mechanism and is used to learn a mutation schedule for different hyperparameters on the fly. Empirical evidence on synthetic functions is provided to show that HPM outperforms hypergradient significantly. Experiments on two benchmark datasets are also conducted to validate the effectiveness of the proposed HPM algorithm for training deep neural networks compared with several strong baselines.

**********

A convex optimization formulation for multivariate regression

Yunzhang Zhu

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

**********

Online Meta-Critic Learning for Off-Policy Actor-Critic Methods

Wei Zhou, Yiying Li, Yongxin Yang, Huaimin Wang, Timothy Hospedales

Off-Policy Actor-Critic (OffP-AC) methods have proven successful in a variety of continuous control tasks. Normally, the critic's action-value function is updated using temporal-difference, and the critic in turn provides a loss for the actor that trains it to take actions with higher expected return. In this paper, we introduce a flexible and augmented meta-critic that observes the learning process and meta-learns an additional loss for the actor that accelerates and improves actor-critic learning. Compared to existing meta-learning algorithms, meta-critic is rapidly learned online for a single task, rather than slowly over a family of tasks. Crucially, our meta-critic is designed for off-policy based learners, which currently provide state-of-the-art reinforcement learning sample efficiency. We demonstrate that online meta-critic learning benefits to a variety of continuous control tasks when combined with contemporary OffP-AC methods DDPG, TD3 and SAC.

**********

The All-or-Nothing Phenomenon in Sparse Tensor PCA

Jonathan Niles-Weed, Ilias Zadik

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

**********

Synthesize, Execute and Debug: Learning to Repair for Neural Program Synthesis

Kavi Gupta, Peter Ebert Christensen, Xinyun Chen, Dawn Song

The use of deep learning techniques has achieved significant progress for program synthesis from input-output examples. However, when the program semantics beco

me more complex, it still remains a challenge to synthesize programs that are co nsistent with the specification. In this work, we propose SED, a neural program generation framework that incorporates synthesis, execution, and debugging stage s. Instead of purely relying on the neural program synthesizer to generate the f inal program, SED first produces initial programs using the neural program synth esizer component, then utilizes a neural program debugger to iteratively repair the generated programs. The integration of the debugger component enables SED to modify the programs based on the execution results and specification, which res embles the coding process of human programmers. On Karel, a challenging input-ou tput program synthesis benchmark, SED reduces the error rate of the neural progr am synthesizer itself by a considerable margin, and outperforms the standard bea m search for decoding.

**********

ARMA Nets: Expanding Receptive Field for Dense Prediction

Jiahao Su, Shiqi Wang, Furong Huang

Global information is essential for dense prediction problems, whose goal is to compute a discrete or continuous label for each pixel in the images. Traditional convolutional layers in neural networks, initially designed for image classific ation, are restrictive in these problems since the filter size limits their rece ptive fields. In this work, we propose to replace any traditional convolutional layer with an autoregressive moving-average (ARMA) layer, a novel module with an adjustable receptive field controlled by the learnable autoregressive coefficie nts. Compared with traditional convolutional layers, our ARMA layer enables expl icit interconnections of the output neurons and learns its receptive field by ad apting the autoregressive coefficients of the interconnections. ARMA layer is ad justable to different types of tasks: for tasks where global information is cruc ial, it is capable of learning relatively large autoregressive coefficients to a llow for an output neuron's receptive field covering the entire input; for tasks where only local information is required, it can learn small or near zero autor egressive coefficients and automatically reduces to a traditional convolutional layer. We show both theoretically and empirically that the effective receptive f ield of networks with ARMA layers (named ARMA networks) expands with larger auto regressive coefficients. We also provably solve the instability problem of learn ing and prediction in the ARMA layer through a re-parameterization mechanism. Ad ditionally, we demonstrate that ARMA networks substantially improve their baseli nes on challenging dense prediction tasks, including video prediction and semant ic segmentation.

**********

Diversity-Guided Multi-Objective Bayesian Optimization With Batch Evaluations

Mina Konakovic Lukovic, Yunsheng Tian, Wojciech Matusik

Many science, engineering, and design optimization problems require balancing th e trade-offs between several conflicting objectives. The objectives are often bl ack-box functions whose evaluations are time-consuming and costly. Multi-objecti ve Bayesian optimization can be used to automate the process of discovering the set of optimal solutions, called Pareto-optimal, while minimizing the number of performed evaluations. To further reduce the evaluation time in the optimization process, testing of several samples in parallel can be deployed. We propose a n ovel multi-objective Bayesian optimization algorithm that iteratively selects th e best batch of samples to be evaluated in parallel. Our algorithm approximates and analyzes a piecewise-continuous Pareto set representation. This representati on allows us to introduce a batch selection strategy that optimizes for both hyp ervolume improvement and diversity of selected samples in order to efficiently a dvance promising regions of the Pareto front. Experiments on both synthetic test functions and real-world benchmark problems show that our algorithm predominant ly outperforms relevant state-of-the-art methods. Code is available at https://g ithub.com/yunshengtian/DGEMO.

**********

SOLOv2: Dynamic and Fast Instance Segmentation

Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, Chunhua Shen

In this work, we design a simple, direct, and fast framework for instance segmen

tation with strong performance. To this end, we propose a novel and effective ap proach, termed SOLOv2, following the principle of the SOLO method [32]. First, o ur new framework is empowered by an efficient and holistic instance mask represe ntation scheme, which dynamically segments each instance in the image, without r esorting to bounding box detection. Specifically, the object mask generation is decoupled into a mask kernel prediction and mask feature learning, which are res ponsible for generating convolution kernels and the feature maps to be convolved with, respectively. Second, SOLOv2 significantly reduces inference overhead wit h our novel matrix non-maximum suppression (NMS) technique. Our Matrix NMS perfo rms NMS with parallel matrix operations in one shot, and yields better results. We demonstrate that the proposed SOLOv2 achieves the state-of-the- art performan ce with high efficiency, making it suitable for both mobile and cloud applicatio ns. A light-weight version of SOLOv2 executes at 31.3 FPS and yields 37.1% AP on COCO test-dev. Moreover, our state-of-the-art results in object detection (from our mask byproduct) and panoptic segmentation show the potential of SOLOv2 to s erve as a new strong baseline for many instance-level recognition tasks. Code is available at https://git.io/AdelaiDet
**********

Robust Recovery via Implicit Bias of Discrepant Learning Rates for Double Over-p arameterization
Chong You, Zhihui Zhu, Qing Qu, Yi Ma
Recent advances have shown that implicit bias of gradient descent on over-parame terized models enables the recovery of low-rank matrices from linear measurement s, even with no prior knowledge on the intrinsic rank. In contrast, for {\em rob ust} low-rank matrix recovery from {\em grossly corrupted} measurements, over-pa rameterization leads to overfitting without prior knowledge on both the intrinsi c rank and sparsity of corruption. This paper shows that with a {\em double over -parameterization} for both the low-rank matrix and sparse corruption, gradient descent with {\em discrepant learning rates} provably recovers the underlying ma trix even without prior knowledge on neither rank of the matrix nor sparsity of the corruption. We further extend our approach for the robust recovery of natura l images by over-parameterizing images with deep convolutional networks. Experim ents show that our method handles different test images and varying corruption l evels with a single learning pipeline where the network width and termination co nditions do not need to be adjusted on a case-by-case basis. Underlying the succ ess is again the implicit bias with discrepant learning rates on different over- parameterized parameters, which may bear on broader applications.
**********

Axioms for Learning from Pairwise Comparisons
Ritesh Noothigattu, Dominik Peters, Ariel D. Procaccia
To be well-behaved, systems that process preference data must satisfy certain co nditions identified by economic decision theory and by social choice theory. In ML, preferences and rankings are commonly learned by fitting a probabilistic mod el to noisy preference data. The behavior of this learning process from the view of economic theory has previously been studied for the case where the data cons ists of rankings. In practice, it is more common to have only pairwise compariso n data, and the formal properties of the associated learning problem are more ch allenging to analyze. We show that a large class of random utility models (inclu ding the Thurstone–Mosteller Model), when estimated using the MLE, satisfy a Par eto efficiency condition. These models also satisfy a strong monotonicity proper ty, which implies that the learning process is responsive to input data. On the other hand, we show that these models fail certain other consistency conditions from social choice theory, and in particular do not always follow the majority o pinion. Our results inform existing and future applications of random utility mo dels for societal decision making.
**********

Continuous Regularized Wasserstein Barycenters
Lingxiao Li, Aude Genevay, Mikhail Yurochkin, Justin M. Solomon
Wasserstein barycenters provide a geometrically meaningful way to aggregate prob ability distributions, built on the theory of optimal transport. They are diffic

ult to compute in practice, however, leading previous work to restrict their sup
ports to finite sets of points. Leveraging a new dual formulation for the regula
rized Wasserstein barycenter problem, we introduce a stochastic algorithm that c
onstructs a continuous approximation of the barycenter. We establish strong dual
ity and use the corresponding primal-dual relationship to parametrize the baryce
nter implicitly using the dual potentials of regularized transport problems. The
 resulting problem can be solved with stochastic gradient descent, which yields
an efficient online algorithm to approximate the barycenter of continuous distri
butions given sample access. We demonstrate the effectiveness of our approach an
d compare against previous work on synthetic examples and real-world application
s.
**********

Spectral Temporal Graph Neural Network for Multivariate Time-series Forecasting
Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai T
ong, Bixiong Xu, Jing Bai, Jie Tong, Qi Zhang
Multivariate time-series forecasting plays a crucial role in many real-world app
lications. It is a challenging problem as one needs to consider both intra-serie
s temporal correlations and inter-series correlations simultaneously. Recently,
there have been multiple works trying to capture both correlations, but most, if
 not all of them only capture temporal correlations in the time domain and resor
t to pre-defined priors as inter-series relationships.
**********

Online Multitask Learning with Long-Term Memory
Mark Herbster, Stephen Pasteris, Lisa Tse
We introduce a novel online multitask setting. In this setting each task is part
itioned into a sequence of segments that is unknown to the learner. Associated w
ith each segment is a hypothesis from some hypothesis class. We give algorithms
that are designed to exploit the scenario where there are many such segments but
 significantly fewer associated hypotheses. We prove regret bounds that hold for
 any segmentation of the tasks and any association of hypotheses to the segments
. In the single-task setting this is equivalent to switching with long-term memo
ry in the sense of [Bousquet and Warmuth 2011]. We provide an algorithm that pre
dicts on each trial in time linear in the number of hypotheses when the hypothes
is class is finite. We also consider infinite hypothesis classes from reproducin
g kernel Hilbert spaces for which we give an algorithm whose per trial time comp
lexity is cubic in the number of cumulative trials. In the single-task special c
ase this is the first example of an efficient regret-bounded switching algorithm
 with long-term memory for a non-parametric hypothesis class.
**********

Fewer is More: A Deep Graph Metric Learning Perspective Using Fewer Proxies
Yuehua Zhu, Muli Yang, Cheng Deng, Wei Liu
Deep metric learning plays a key role in various machine learning tasks. Most of
 the previous works have been confined to sampling from a mini-batch, which cann
ot precisely characterize the global geometry of the embedding space. Although r
esearchers have developed proxy- and classification-based methods to tackle the
sampling issue, those methods inevitably incur a redundant computational cost. I
n this paper, we propose a novel Proxy-based deep Graph Metric Learning (ProxyGM
L) approach from the perspective of graph classification, which uses fewer proxi
es yet achieves better comprehensive performance. Specifically, multiple global
proxies are leveraged to collectively approximate the original data points for e
ach class. To efficiently capture local neighbor relationships, a small number o
f such proxies are adaptively selected to construct similarity subgraphs between
 these proxies and each data point. Further, we design a novel reverse label pro
pagation algorithm, by which the neighbor relationships are adjusted according t
o ground-truth labels, so that a discriminative metric space can be learned duri
ng the process of subgraph classification. Extensive experiments carried out on
widely-used CUB-200-2011, Cars196, and Stanford Online Products datasets demonst
rate the superiority of the proposed ProxyGML over the state-of-the-art methods
in terms of both effectiveness and efficiency. The source code is publicly avail
able at \url{https://github.com/YuehuaZhu/ProxyGML}.

**********
Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting
LEI BAI, Lina Yao, Can Li, Xianzhi Wang, Can Wang
Modeling complex spatial and temporal correlations in the correlated time series data is indispensable for understanding the traffic dynamics and predicting the future status of an evolving traffic system. Recent works focus on designing complicated graph neural network architectures to capture shared patterns with the help of pre-defined graphs. In this paper, we argue that learning node-specific patterns is essential for traffic forecasting while pre-defined graph is avoidable.
To this end, we propose two adaptive modules for enhancing Graph Convolutional Network (GCN) with new capabilities: 1) a Node Adaptive Parameter Learning (NAPL) module to capture node-specific patterns; 2) a Data Adaptive Graph Generation (DAGG) module to infer the inter-dependencies among different traffic series automatically. We further propose an Adaptive Graph Convolutional Recurrent Network (AGCRN) to capture fine-grained spatial and temporal correlations in traffic series automatically based on the two modules and recurrent networks. Our experiments on two real-world traffic datasets show AGCRN outperforms state-of-the-art by a significant margin without pre-defined graphs about spatial connections.
**********
On Reward-Free Reinforcement Learning with Linear Function Approximation
Ruosong Wang, Simon S. Du, Lin Yang, Russ R. Salakhutdinov
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
Robustness of Community Detection to Random Geometric Perturbations
Sandrine Peche, Vianney Perchet
We consider the stochastic block model where connection between vertices is perturbed by some latent (and unobserved) random geometric graph. The objective is to prove that spectral methods are robust to this type of noise, even if they are agnostic to the presence (or not) of the random graph. We provide explicit regimes where the second eigenvector of the adjacency matrix is highly correlated to the true community vector (and therefore when weak/exact recovery is possible). This is possible thanks to a detailed analysis of the spectrum of the latent random graph, of its own interest.
**********
Learning outside the Black-Box: The pursuit of interpretable models
Jonathan Crabbe, Yao Zhang, William Zame, Mihaela van der Schaar
Machine learning has proved its ability to produce accurate models -- but the deployment of these models outside the machine learning community has been hindered by the difficulties of interpreting these models. This paper proposes an algorithm that produces a continuous global interpretation of any given continuous black-box function. Our algorithm employs a variation of projection pursuit in which the ridge functions are chosen to be Meijer G-functions, rather than the usual polynomial splines. Because Meijer G-functions are differentiable in their parameters, we can "tune" the parameters of the representation by gradient descent; as a consequence, our algorithm is efficient. Using five familiar data sets from the UCI repository and two familiar machine learning algorithms, we demonstrate that our algorithm produces global interpretations that are both faithful (highly accurate) and parsimonious (involve a small number of terms). Our interpretations permit easy understanding of the relative importance of features and feature interactions. Our interpretation algorithm represents a leap forward from the previous state of the art.
**********
Breaking Reversibility Accelerates Langevin Dynamics for Non-Convex Optimization
Xuefeng GAO, Mert Gurbuzbalaban, Lingjiong Zhu
Langevin dynamics (LD) has been proven to be a powerful technique for optimizing a non-convex objective as an efficient algorithm to find local minima while eve

ntually visiting a global minimum on longer time-scales. LD is based on the firs t-order Langevin diffusion which is reversible in time. We study two variants th at are based on non-reversible Langevin diffusions: the underdamped Langevin dyn amics (ULD) and the Langevin dynamics with a non-symmetric drift (NLD). Adopting the techniques of Tzen et al. (2018) for LD to non-reversible diffusions, we sh ow that for a given local minimum that is within an arbitrary distance from the initialization, with high probability, either the ULD trajectory ends up somewhe re outside a small neighborhood of this local minimum within a recurrence time w hich depends on the smallest eigenvalue of the Hessian at the local minimum or t hey enter this neighborhood by the recurrence time and stay there for a potentia lly exponentially long escape time. The ULD algorithm improves upon the recurren ce time obtained for LD in Tzen et al. (2018) with respect to the dependency on the smallest eigenvalue of the Hessian at the local minimum. Similar results and improvements are obtained for the NLD algorithm. We also show that non-reversib le variants can exit the basin of attraction of a local minimum faster in discre te time when the objective has two local minima separated by a saddle point and quantify the amount of improvement. Our analysis suggests that non-reversible La ngevin algorithms are more efficient to locate a local minimum as well as explor ing the state space.
**********

Robust large-margin learning in hyperbolic space
Melanie Weber, Manzil Zaheer, Ankit Singh Rawat, Aditya K. Menon, Sanjiv Kumar
Recently, there has been a surge of interest in representation learning in hyper bolic spaces, driven by their ability to represent hierarchical data with signif icantly fewer dimensions than standard Euclidean spaces. However, the viability and benefits of hyperbolic spaces for downstream machine learning tasks have rec eived less attention. In this paper, we present, to our knowledge, the first the oretical guarantees for learning a classifier in hyperbolic rather than Euclidea n space. Specifically, we consider the problem of learning a large-margin classi fier for data possessing a hierarchical structure. Our first contribution is a h yperbolic perceptron algorithm, which provably converges to a separating hyperpl ane. We then provide an algorithm to efficiently learn a large-margin hyperplane , relying on the careful injection of adversarial examples. Finally, we prove th at for hierarchical data that embeds well into hyperbolic space, the low embeddi ng dimension ensures superior guarantees when learning the classifier directly i n hyperbolic space.
**********

Replica-Exchange Nos\'e-Hoover Dynamics for Bayesian Learning on Large Datasets
Rui Luo, Qiang Zhang, Yaodong Yang, Jun Wang
In this paper, we present a new practical method for Bayesian learning that can rapidly draw representative samples from complex posterior distributions with mu ltiple isolated modes in the presence of mini-batch noise.
This is achieved by simulating a collection of replicas in parallel with differe nt temperatures and periodically swapping them.
When evolving the replicas' states, the Nos\'e-Hoover dynamics is applied, which adaptively neutralizes the mini-batch noise.
To perform proper exchanges, a new protocol is developed with a noise-aware test of acceptance, by which the detailed balance is reserved in an asymptotic way.
While its efficacy on complex multimodal posteriors has been illustrated by test ing over synthetic distributions, experiments with deep Bayesian neural networks on large-scale datasets have shown its significant improvements over strong bas elines.
**********

Adversarially Robust Few-Shot Learning: A Meta-Learning Approach
Micah Goldblum, Liam Fowl, Tom Goldstein
Previous work on adversarially robust neural networks for image classification r equires large training sets and computationally expensive training procedures.
On the other hand, few-shot learning methods are highly vulnerable to adversaria l examples.  The goal of our work is to produce networks which both perform well at few-shot classification tasks and are simultaneously robust to adversarial e

xamples.  We develop an algorithm, called Adversarial Querying (AQ), for produci
ng adversarially robust meta-learners, and we thoroughly investigate the causes
for adversarial vulnerability.  Moreover, our method achieves far superior robus
t performance on few-shot image classification tasks, such as Mini-ImageNet and
CIFAR-FS, than robust transfer learning.
**********

Neural Anisotropy Directions
Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi, Pascal Frossard
In this work, we analyze the role of the network architecture in shaping the ind
uctive bias of deep classifiers. To that end, we start by focusing on a very sim
ple problem, i.e., classifying a class of linearly separable distributions, and
show that, depending on the direction of the discriminative feature of the distr
ibution, many state-of-the-art deep convolutional neural networks (CNNs) have a
surprisingly hard time solving this simple task. We then define as neural anisot
ropy directions (NADs) the vectors that encapsulate the directional inductive bi
as of an architecture. These vectors, which are specific for each architecture a
nd hence act as a signature, encode the preference of a network to separate the
input data based on some particular features. We provide an efficient method to
identify NADs for several CNN architectures and thus reveal their directional in
ductive biases. Furthermore, we show that, for the CIFAR-10 dataset, NADs charac
terize the features used by CNNs to discriminate between different classes.
**********

Digraph Inception Convolutional Networks
Zekun Tong, Yuxuan Liang, Changsheng Sun, Xinke Li, David Rosenblum, Andrew Lim
Graph Convolutional Networks (GCNs) have shown promising results in modeling gra
ph-structured data. However, they have difficulty with processing digraphs becau
se of two reasons: 1) transforming directed to undirected graph to guarantee the
 symmetry of graph Laplacian is not reasonable since it not only misleads messag
e passing scheme to aggregate incorrect weights but also deprives the unique cha
racteristics of digraph structure; 2) due to the fixed receptive field in each l
ayer, GCNs fail to obtain multi-scale features that can boost their performance.
 In this paper, we theoretically extend spectral-based graph convolution to digr
aphs and derive a simplified form using personalized PageRank. Specifically, we
present the Digraph Inception Convolutional Networks (DiGCN) which utilizes digr
aph convolution and kth-order proximity to achieve larger receptive fields and l
earn multi-scale features in digraphs. We empirically show that DiGCN can encode
 more structural information from digraphs than GCNs and help achieve better per
formance when generalized to other models. Moreover, experiments on various benc
hmarks demonstrate its superiority against the state-of-the-art methods.
**********

PAC-Bayesian Bound for the Conditional Value at Risk
Zakaria Mhammedi, Benjamin Guedj, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Stochastic Stein Discrepancies
Jackson Gorham, Anant Raj, Lester Mackey
Stein discrepancies (SDs) monitor convergence and non-convergence in approximate
 inference when exact integration and sampling are intractable. However, the com
putation of a Stein discrepancy can be prohibitive if the Stein operator -- ofte
n a sum over likelihood terms or potentials -- is expensive to evaluate. To addr
ess this deficiency, we show that stochastic Stein discrepancies (SSDs) based on
 subsampled approximations of the Stein operator inherit the convergence control
 properties of standard SDs with probability 1. Along the way, we establish the
convergence of Stein variational gradient descent (SVGD) on unbounded domains, r
esolving an open question of Liu (2017). In our experiments with biased Markov c
hain Monte Carlo (MCMC) hyperparameter tuning, approximate MCMC sampler selectio
n, and stochastic SVGD, SSDs deliver comparable inferences to standard SDs with

orders of magnitude fewer likelihood evaluations.
**********
On the Role of Sparsity and DAG Constraints for Learning Linear DAGs
Ignavier Ng, AmirEmad Ghassami, Kun Zhang
Learning graphical structure based on Directed Acyclic Graphs (DAGs) is a challenging problem, partly owing to the large search space of possible graphs. A recent line of work formulates the structure learning problem as a continuous constrained optimization task using the least squares objective and an algebraic characterization of DAGs. However, the formulation requires a hard DAG constraint and may lead to optimization difficulties. In this paper, we study the asymptotic role of the sparsity and DAG constraints for learning DAG models in the linear Gaussian and non-Gaussian cases, and investigate their usefulness in the finite sample regime. Based on the theoretical results, we formulate a likelihood-based score function, and show that one only has to apply soft sparsity and DAG constraints to learn a DAG equivalent to the ground truth DAG. This leads to an unconstrained optimization problem that is much easier to solve. Using gradient-based optimization and GPU acceleration, our procedure can easily handle thousands of nodes while retaining a high accuracy. Extensive experiments validate the effectiveness of our proposed method and show that the DAG-penalized likelihood objective is indeed favorable over the least squares one with the hard DAG constraint.
**********
Cream of the Crop: Distilling Prioritized Paths For One-Shot Neural Architecture Search
Houwen Peng, Hao Du, Hongyuan Yu, QI LI, Jing Liao, Jianlong Fu
One-shot weight sharing methods have recently drawn great attention in neural architecture search due to high efficiency and competitive performance. However, weight sharing across models has an inherent deficiency, i.e., insufficient training of subnetworks in the hypernetwork. To alleviate this problem, we present a simple yet effective architecture distillation method. The central idea is that subnetworks can learn collaboratively and teach each other throughout the training process, aiming to boost the convergence of individual models. We introduce the concept of prioritized path, which refers to the architecture candidates exhibiting superior performance during training. Distilling knowledge from the prioritized paths is able to boost the training of subnetworks. Since the prioritized paths are changed on the fly depending on their performance and complexity, the final obtained paths are the cream of the crop. We directly select the most promising one from the prioritized paths as the final architecture, without using other complex search methods, such as reinforcement learning or evolution algorithms. The experiments on ImageNet verify such path distillation method can improve the convergence ratio and performance of the hypernetwork, as well as boosting the training of subnetworks. The discovered architectures achieve superior performance compared to the recent MobileNetV3 and EfficientNet families under aligned settings. Moreover, the experiments on object detection and more challenging search space show the generality and robustness of the proposed method. Code and models are available at \url{https://github.com/neurips-20/cream.git}.
**********
Fair Multiple Decision Making Through Soft Interventions
Yaowei Hu, Yongkai Wu, Lu Zhang, Xintao Wu
Previous research in fair classification mostly focuses on a single decision model. In reality, there usually exist multiple decision models within a system and all of which may contain a certain amount of discrimination. Such realistic scenarios introduce new challenges to fair classification: since discrimination may be transmitted from upstream models to downstream models, building decision models separately without taking upstream models into consideration cannot guarantee to achieve fairness. In this paper, we propose an approach that learns multiple classifiers and achieves fairness for all of them simultaneously, by treating each decision model as a soft intervention and inferring the post-intervention distributions to formulate the loss function as well as the fairness constraints. We adopt surrogate functions to smooth the loss function and constraints, and theoretically show that the excess risk of the proposed loss function can be boun

ded in a form that is the same as that for traditional surrogated loss functions . Experiments using both synthetic and real-world datasets show the effectivenes s of our approach.
**********

Representation Learning for Integrating Multi-domain Outcomes to Optimize Indivi dualized Treatment
Yuan Chen, Donglin Zeng, Tianchen Xu, Yuanjia Wang
For mental disorders, patients' underlying mental states are non-observed latent constructs which have to be inferred from observed multi-domain measurements su ch as diagnostic symptoms and patient functioning scores. Additionally, substant ial heterogeneity in the disease diagnosis between patients needs to be addresse d for optimizing individualized treatment policy in order to achieve precision m edicine. To address these challenges, we propose an integrated learning framewor k that can simultaneously learn patients' underlying mental states and recommend optimal treatments for each individual. This learning framework is based on the measurement theory in psychiatry for modeling multiple disease diagnostic measu res as arising from the underlying causes (true mental states). It allows incorp oration of the multivariate pre- and post-treatment outcomes as well as biologic al measures while preserving the invariant structure for representing patients' latent mental states. A multi-layer neural network is used to allow complex trea tment effect heterogeneity. Optimal treatment policy can be inferred for future patients by comparing their potential mental states under different treatments g iven the observed multi-domain pre-treatment measurements. Experiments on simula ted data and a real-world clinical trial data show that the learned treatment po lices compare favorably to alternative methods on heterogeneous treatment effect s, and have broad utilities which lead to better patient outcomes on multiple do mains.
**********

Learning to Play No-Press Diplomacy with Best Response Policy Iteration
Thomas Anthony, Tom Eccles, Andrea Tacchetti, János Kramár, Ian Gemp, Thomas Hud son, Nicolas Porcel, Marc Lanctot, Julien Perolat, Richard Everett, Satinder Sin gh, Thore Graepel, Yoram Bachrach
Recent advances in deep reinforcement learning (RL) have led to considerable pro gress in many 2-player zero-sum games, such as Go, Poker and Starcraft. The pure ly adversarial nature of such games allows for conceptually simple and principle d application of RL methods. However real-world settings are many-agent, and age nt interactions are complex mixtures of common-interest and competitive aspects. We consider Diplomacy, a 7-player board game designed to accentuate dilemmas re sulting from many-agent interactions. It also features a large combinatorial act ion space and simultaneous moves, which are challenging for RL algorithms. We pr opose a simple yet effective approximate best response operator, designed to han dle large combinatorial action spaces and simultaneous moves. We also introduce a family of policy iteration methods that approximate fictitious play. With thes e methods, we successfully apply RL to Diplomacy: we show that our agents convin cingly outperform the previous state-of-the-art, and game theoretic equilibrium analysis shows that the new process yields consistent improvements.
**********

Inverse Learning of Symmetries
Mario Wieser, Sonali Parbhoo, Aleksander Wieczorek, Volker Roth
Symmetry transformations induce invariances and are a crucial building block of modern machine learning algorithms. In many complex domains, such as the chemica l space, invariances can be observed, yet the corresponding symmetry transformat ion cannot be formulated analytically. We propose to learn the symmetry transfor mation with a model consisting of two latent subspaces, where the first subspace captures the target and the second subspace the remaining invariant information . Our approach is based on the deep information bottleneck in combination with a continuous mutual information regulariser. Unlike previous methods, we focus on the challenging task of minimising mutual information in continuous domains. To this end, we base the calculation of mutual information on correlation matrices in combination with a bijective variable transformation. Extensive experiments

demonstrate that our model outperforms state-of-the-art methods on artificial an
d molecular datasets.
**********
DiffGCN: Graph Convolutional Networks via Differential Operators and Algebraic M
ultigrid Pooling
Moshe Eliasof, Eran Treister
Graph Convolutional Networks (GCNs) have shown to be effective in handling
unordered data like point clouds and meshes. In this work we propose novel
approaches for graph convolution, pooling and unpooling, inspired from finite
differences and algebraic multigrid frameworks. We form a parameterized convolut
ion
kernel based on discretized differential operators, leveraging the graph mass,
gradient and Laplacian. This way, the parameterization does not depend on the
graph structure, only on the meaning of the network convolutions as differential
operators. To allow hierarchical representations of the input, we propose poolin
g
and unpooling operations that are based on algebraic multigrid methods, which
are mainly used to solve partial differential equations on unstructured grids. T
o
motivate and explain our method, we compare it to standard convolutional neural
networks, and show their similarities and relations in the case of a regular gri
d. Our
proposed method is demonstrated in various experiments like classification and
part-segmentation, achieving on par or better than state of the art results. We
also
analyze the computational cost of our method compared to other GCNs.
**********
Distributed Newton Can Communicate Less and Resist Byzantine Workers
Avishek Ghosh, Raj Kumar Maity, Arya Mazumdar
We develop a distributed second order optimization algorithm that is communicati
on-efficient as well as robust against Byzantine failures of the worker machines
. We propose an iterative approximate Newton-type algorithm, where the worker ma
chines communicate \emph{only once} per iteration with the central machine. This
 is in sharp contrast with the state-of-the-art distributed second order algorit
hms like GIANT \cite{giant}, DINGO\cite{dingo}, where the worker machines send (
functions of) local gradient and Hessian sequentially; thus ending up communicat
ing twice with the central machine per iteration. Furthermore, we employ a simpl
e norm based thresholding rule to filter-out the Byzantine worker machines. We e
stablish the linear-quadratic rate of convergence of our proposed algorithm and
establish that the communication savings and Byzantine resilience attributes onl
y correspond to a small statistical error rate for arbitrary convex loss functio
ns. To the best of our knowledge, this is the first work that addresses the issu
e of Byzantine resilience in second order distributed optimization. Furthermore,
 we validate our theoretical results with extensive experiments on synthetically
 generated and benchmark LIBSVM \cite{libsvm} data-set and demonstrate convergen
ce guarantees.
**********
Efficient Nonmyopic Bayesian Optimization via One-Shot Multi-Step Trees
Shali Jiang, Daniel Jiang, Maximilian Balandat, Brian Karrer, Jacob Gardner, Rom
an Garnett
Bayesian optimization is a sequential decision making framework for optimizing e
xpensive-to-evaluate black-box functions. Computing a full lookahead policy amou
nts to solving a highly intractable stochastic dynamic program. Myopic approache
s, such as expected improvement, are often adopted in practice, but they ignore
the long-term impact of the immediate decision. Existing nonmyopic approaches ar
e mostly heuristic and/or computationally expensive. In this paper, we provide t
he first efficient implementation of general multi-step lookahead Bayesian optim
ization, formulated as a sequence of nested optimization problems within a multi
-step scenario tree. Instead of solving these problems in a nested way, we equiv
alently optimize all decision variables in the full tree jointly, in a "one-shot

" fashion. Combining this with an efficient method for implementing multi-step G
aussian process "fantasization," we demonstrate that multi-step expected improve
ment is computationally tractable and exhibits performance superior to existing
methods on a wide range of benchmarks.
**********
Effective Diversity in Population Based Reinforcement Learning
Jack Parker-Holder, Aldo Pacchiano, Krzysztof M. Choromanski, Stephen J. Roberts
Exploration is a key problem in reinforcement learning, since agents can only le
arn from data they acquire in the environment. With that in mind, maintaining a
population of agents is an attractive method, as it allows data be collected wit
h a diverse set of behaviors. This behavioral diversity is often boosted via mul
ti-objective loss functions. However, those approaches typically leverage mean f
ield updates based on pairwise distances, which makes them susceptible to cyclin
g behaviors and increased redundancy. In addition, explicitly boosting diversity
often has a detrimental impact on optimizing already fruitful behaviors for rew
ards. As such, the reward-diversity trade off typically relies on heuristics. Fi
nally, such methods require behavioral representations, often handcrafted and do
main specific. In this paper, we introduce an approach to optimize all members o
f a population simultaneously. Rather than using pairwise distance, we measure t
he volume of the entire population in a behavioral manifold, defined by task-agn
ostic behavioral embeddings. In addition, our algorithm Diversity via Determinan
ts (DvD), adapts the degree of diversity during training using online learning t
echniques. We introduce both evolutionary and gradient-based instantiations of D
vD and show they effectively improve exploration without reducing performance wh
en better exploration is not required.
**********
Elastic-InfoGAN: Unsupervised Disentangled Representation Learning in Class-Imba
lanced Data
Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, Yong Jae Lee
We propose a novel unsupervised generative model that learns to disentangle obje
ct identity from other low-level aspects in class-imbalanced data. We first inve
stigate the issues surrounding the assumptions about uniformity made by InfoGAN,
and demonstrate its ineffectiveness to properly disentangle object identity in
imbalanced data. Our key idea is to make the discovery of the discrete latent fa
ctor of variation invariant to identity-preserving transformations in real image
s, and use that as a signal to learn the appropriate latent distribution represe
nting object identity. Experiments on both artificial (MNIST, 3D cars, 3D chairs
, ShapeNet) and real-world (YouTube-Faces) imbalanced datasets demonstrate the e
ffectiveness of our method in disentangling object identity as a latent factor o
f variation.
**********
Direct Policy Gradients: Direct Optimization of Policies in Discrete Action Spac
es
Guy Lorberbom, Chris J. Maddison, Nicolas Heess, Tamir Hazan, Daniel Tarlow
Direct optimization (McAllester et al., 2010; Song et al., 2016) is an appealing
framework that replaces integration with optimization of a random objective for
approximating gradients in models with discrete random variables (Lorberbom et
al., 2018). A* sampling (Maddison et al., 2014) is a framework for optimizing s
uch random objectives over large spaces. We show how to combine these technique
s to yield a reinforcement learning algorithm that approximates a policy gradien
t by finding trajectories that optimize a random objective. We call the resulti
ng algorithms \emph{direct policy gradient} (DirPG) algorithms. A main benefit o
f DirPG algorithms is that they allow the insertion of domain knowledge in the f
orm of upper bounds on return-to-go at training time, like is used in heuristic
search, while still directly computing a policy gradient. We further analyze the
ir properties, showing there are cases where DirPG has an exponentially larger p
robability of sampling informative gradients compared to REINFORCE. We also show
that there is a built-in variance reduction technique and that a parameter that
was previously viewed as a numerical approximation can be interpreted as contro
lling risk sensitivity. Empirically, we evaluate the effect of key degrees of fr

eedom and show that the algorithm performs well in illustrative domains compared to baselines.

**********

Hybrid Models for Learning to Branch

Prateek Gupta, Maxime Gasse, Elias Khalil, Pawan Mudigonda, Andrea Lodi, Yoshua Bengio

A recent Graph Neural Network (GNN) approach for learning to branch has been shown to successfully reduce the running time of branch-and-bound algorithms for Mixed Integer Linear Programming (MILP). While the GNN relies on a GPU for inference, MILP solvers are purely CPU-based. This severely limits its application as many practitioners may not have access to high-end GPUs. In this work, we ask two key questions. First, in a more realistic setting where only a CPU is available, is the GNN model still competitive? Second, can we devise an alternate computationally inexpensive model that retains the predictive power of the GNN architecture? We answer the first question in the negative, and address the second question by proposing a new hybrid architecture for efficient branching on CPU machines. The proposed architecture combines the expressive power of GNNs with computationally inexpensive multi-layer perceptrons (MLP) for branching. We evaluate our methods on four classes of MILP problems, and show that they lead to up to 26% reduction in solver running time compared to state-of-the-art methods without a GPU, while extrapolating to harder problems than it was trained on. The code for this project is publicly available at https://github.com/pg2455/Hybrid-learn2branch.

**********

WoodFisher: Efficient Second-Order Approximation for Neural Network Compression

Sidak Pal Singh, Dan Alistarh

Second-order information, in the form of Hessian- or Inverse-Hessian-vector products, is a fundamental tool for solving optimization problems. Recently, there has been significant interest in utilizing this information in the context of deep neural networks; however, relatively little is known about the quality of existing approximations in this context. Our work considers this question, examines the accuracy of existing approaches, and proposes a method called WoodFisher to compute a faithful and efficient estimate of the inverse Hessian.

**********

Bi-level Score Matching for Learning Energy-based Latent Variable Models

Fan Bao, Chongxuan LI, Kun Xu, Hang Su, Jun Zhu, Bo Zhang

Score matching (SM) provides a compelling approach to learn energy-based models (EBMs) by avoiding the calculation of partition function. However, it remains largely open to learn energy-based latent variable models (EBLVMs), except some special cases. This paper presents a bi-level score matching (BiSM) method to learn EBLVMs with general structures by reformulating SM as a bi-level optimization problem. The higher level introduces a variational posterior of the latent variables and optimizes a modified SM objective, and the lower level optimizes the variational posterior to fit the true posterior. To solve BiSM efficiently, we develop a stochastic optimization algorithm with gradient unrolling. Theoretically, we analyze the consistency of BiSM and the convergence of the stochastic algorithm. Empirically, we show the promise of BiSM in Gaussian restricted Boltzmann machines and highly nonstructural EBLVMs parameterized by deep convolutional neural networks. BiSM is comparable to the widely adopted contrastive divergence and SM methods when they are applicable; and can learn complex EBLVMs with intractable posteriors to generate natural images.

**********

Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding

Zhu Zhang, Zhou Zhao, Zhijie Lin, jieming zhu, Xiuqiang He

Weakly-supervised vision-language grounding aims to localize a target moment in a video or a specific region in an image according to the given sentence query, where only video-level or image-level sentence annotations are provided during training. Most existing approaches employ the MIL-based or reconstruction-based paradigms for the WSVLG task, but the former heavily depends on the quality of ra

ndomly-selected negative samples and the latter cannot directly optimize the vis
ual-textual alignment score. In this paper, we propose a novel Counterfactual Co
ntrastive Learning (CCL) to develop sufficient contrastive training between coun
terfactual positive and negative results, which are based on robust and destruct
ive counterfactual transformations. Concretely, we design three counterfactual t
ransformation strategies from the feature-, interaction- and relation-level, whe
re the feature-level method damages the visual features of selected proposals, i
nteraction-level approach confuses the vision-language interaction and relation-
level strategy destroys the context clues in proposal relationships. Extensive e
xperiments on five vision-language grounding datasets verify the effectiveness o
f our CCL paradigm.
**********

Decision trees as partitioning machines to characterize their generalization pro
perties
Jean-Samuel Leboeuf, Frédéric LeBlanc, Mario Marchand
Requests for name changes in the electronic proceedings will be accepted with 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 to Prove Theorems by Learning to Generate Theorems
Mingzhe Wang, Jia Deng
We consider the task of automated theorem proving, a key AI task. Deep learning
 has shown promise for training theorem provers, but there are limited human-writ
ten theorems and proofs available for supervised learning. To address this limit
ation, we propose to learn a neural generator that automatically synthesizes the
orems and proofs for the purpose of training a theorem prover. Experiments on re
al-world  tasks demonstrate that synthetic data from our approach  improves the
theorem prover and advances the state of the art of automated theorem proving in
 Metamath.
**********

3D Self-Supervised Methods for Medical Imaging
Aiham Taleb, Winfried Loetzsch, Noel  Danz, Julius Severin, Thomas Gaertner, Ben
jamin Bergner, Christoph Lippert
Self-supervised learning methods have witnessed a recent surge of interest after
 proving successful in multiple application fields.
In this work, we leverage these techniques, and we propose 3D versions for five
different self-supervised methods, in the form of proxy tasks. Our methods facil
itate neural network feature learning from unlabeled 3D images, aiming to reduce
 the required cost for expert annotation.
The developed algorithms are 3D Contrastive Predictive Coding, 3D Rotation predi
ction, 3D Jigsaw puzzles, Relative 3D patch location, and 3D Exemplar networks.
Our experiments show that pretraining models with our 3D tasks yields more power
ful semantic representations, and enables solving downstream tasks more accurate
ly and efficiently, compared to training the models from scratch and to pretrain
ing them on 2D slices.
We demonstrate the effectiveness of our methods on three downstream tasks from t
he medical imaging domain: i) Brain Tumor Segmentation from 3D MRI, ii) Pancreas
 Tumor Segmentation from 3D CT, and iii) Diabetic Retinopathy Detection from 2D
Fundus images. In each task, we assess the gains in data-efficiency, performance
, and speed of convergence. Interestingly, we also find gains when transferring
the learned representations, by our methods, from a large unlabeled 3D corpus to
 a small downstream-specific dataset.
We achieve results competitive to state-of-the-art solutions at a fraction of th
e computational expense.
We publish our implementations for the developed algorithms (both 3D and 2D vers
ions) as an open-source library, in an effort to allow other researchers to appl
y and extend our methods on their datasets.
**********

 Bayesian filtering unifies adaptive and non-adaptive neural network optimizatio

n methods
Laurence Aitchison
We formulate the problem of neural network optimization as Bayesian filtering, w
here the observations are backpropagated gradients.  While neural network optimi
zation has previously been studied using natural gradient methods which are clos
ely related to Bayesian inference, they were unable to recover standard optimize
rs such as Adam and RMSprop with a root-mean-square gradient normalizer, instead
 getting a mean-square normalizer.  To recover the root-mean-square normalizer,
we find it necessary to account for the temporal dynamics of all the other param
eters as they are optimized.  The resulting optimizer, AdaBayes, adaptively tran
sitions between SGD-like and Adam-like behaviour, automatically recovers AdamW,
a state of the art variant of Adam with decoupled weight decay, and has generali
sation performance competitive with SGD.
**********

Worst-Case Analysis for Randomly Collected Data
Justin Chen, Gregory Valiant, Paul Valiant
We introduce a framework for statistical estimation that leverages knowledge of
how samples are collected but makes no distributional assumptions on the data va
lues. Specifically, we consider a population of elements [n]={1,...,n} with corr
esponding data values x1,...,xn. We observe the values for a "sample" set A \sub
set [n] and wish to estimate some statistic of the values for a "target" set B \
subset [n] where B could be the entire set. Crucially, we assume that the sets A
 and B are drawn according to some known distribution P over pairs of subsets of
 [n]. A given estimation algorithm is evaluated based on its "worst-case, expect
ed error" where the expectation is with respect to the distribution P from which
 the sample A and target sets B are drawn, and the worst-case is with respect to
 the data values x1,...,xn. Within this framework, we give an efficient algorith
m for estimating the target mean that returns a weighted combination of the samp
le values--where the weights are functions of the distribution P and the sample
and target sets A, B--and show that the worst-case expected error achieved by th
is algorithm is at most a multiplicative pi/2 factor worse than the optimal of s
uch algorithms.  The algorithm and proof leverage a surprising connection to the
 Grothendieck problem. We also extend these results to the linear regression set
ting where each datapoint is not a scalar but a labeled vector (xi,yi). This fra
mework, which makes no distributional assumptions on the data values but rather
relies on knowledge of the data collection process via the distribution P, is a
significant departure from the typical statistical estimation framework and intr
oduces a uniform analysis for the many natural settings where membership in a sa
mple may be correlated with data values, such as when individuals are recruited
into a sample through their social networks as in "snowball/chain" sampling  or
when samples have chronological structure as in "selective prediction".
**********

Truthful Data Acquisition via Peer Prediction
Yiling Chen, Yiheng Shen, Shuran Zheng
We consider the problem of purchasing data for machine learning or statistical e
stimation. The data analyst has a budget to purchase datasets from multiple data
 providers. She does not have any test data that can be used to evaluate the col
lected data and can assign payments to data providers solely based on the collec
ted datasets. We consider the problem in the standard Bayesian paradigm and in t
wo settings: (1) data are only collected once; (2) data are collected repeatedly
 and each day's data are drawn independently from the same distribution. For bot
h settings, our mechanisms guarantee that truthfully reporting one's dataset is
always an equilibrium by adopting techniques from peer prediction: pay each prov
ider the mutual information between his reported data and other providers' repor
ted data. Depending on the data distribution, the mechanisms can also discourage
 misreports that would lead to inaccurate predictions. Our mechanisms also guara
ntee individual rationality and budget feasibility for certain underlying distri
butions in the first setting and for all distributions in the second setting.
**********

Learning Robust Decision Policies from Observational Data

Muhammad Osama, Dave Zachariah, Peter Stoica
We address the problem of learning a decision policy from observational data of past decisions in contexts with features and associated outcomes. The past policy maybe unknown and in safety-critical applications, such as medical decision support, it is of interest to learn robust policies that reduce the risk of outcomes with high costs. In this paper, we develop a method for learning policies that reduce tails of the cost distribution at a specified level and, moreover, provide a statistically valid bound on the cost of each decision. These properties are valid under finite samples -- even in scenarios with uneven or no overlap between features for different decisions in the observed data -- by building on recent results in conformal prediction. The performance and statistical properties of the proposed method are illustrated using both real and synthetic data.
**********

Byzantine Resilient Distributed Multi-Task Learning
Jiani Li, Waseem Abbas, Xenofon Koutsoukos
Distributed multi-task learning provides significant advantages in multi-agent networks with heterogeneous data sources where agents aim to learn distinct but correlated models simultaneously. However, distributed algorithms for learning relatedness among tasks are not resilient in the presence of Byzantine agents. In this paper, we present an approach for Byzantine resilient distributed multi-task learning. We propose an efficient online weight assignment rule by measuring the accumulated loss using an agent’s data and its neighbors’ models. A small accumulated loss indicates a large similarity between the two tasks. In order to ensure the Byzantine resilience of the aggregation at a normal agent, we introduce a step for filtering out larger losses. We analyze the approach for convex models and show that normal agents converge resiliently towards their true targets. Further, an agent’s learning performance using the proposed weight assignment rule is guaranteed to be at least as good as in the non-cooperative case as measured by the expected regret. Finally, we demonstrate the approach using three case studies, including regression and classification problems, and show that our method exhibits good empirical performance for non-convex models, such as convolutional neural networks.
**********

Reinforcement Learning in Factored MDPs: Oracle-Efficient Algorithms and Tighter Regret Bounds for the Non-Episodic Setting
Ziping Xu, 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.
**********

Improving model calibration with accuracy versus uncertainty optimization
Ranganath Krishnan, Omesh Tickoo
Obtaining reliable and accurate quantification of uncertainty estimates from deep neural networks is important in safety-critical applications. A well-calibrated model should be accurate when it is certain about its prediction and indicate high uncertainty when it is likely to be inaccurate. Uncertainty calibration is a challenging problem as there is no ground truth available for uncertainty estimates. We propose an optimization method that leverages the relationship between accuracy and uncertainty as an anchor for uncertainty calibration. We introduce a differentiable accuracy versus uncertainty calibration (AvUC) loss function that allows a model to learn to provide well-calibrated uncertainties, in addition to improved accuracy. We also demonstrate the same methodology can be extended to post-hoc uncertainty calibration on pretrained models. We illustrate our approach with mean-field stochastic variational inference and compare with state-of-the-art methods. Extensive experiments demonstrate our approach yields better model calibration than existing methods on large-scale image classification tasks under distributional shift.
**********

The Convolution Exponential and Generalized Sylvester Flows

Emiel Hoogeboom, Victor Garcia Satorras, Jakub Tomczak, Max Welling
This paper introduces a new method to build linear flows, by taking the exponent ial of a linear transformation. This linear transformation does not need to be i nvertible itself, and the exponential has the following desirable properties: it  is guaranteed to be invertible, its inverse is straightforward to compute and t he log Jacobian determinant is equal to the trace of the linear transformation.  An important insight is that the exponential can be computed implicitly, which a llows the use of convolutional layers. Using this insight, we develop new invert ible transformations named convolution exponentials and graph convolution expone ntials, which retain the equivariance of their underlying transformations. In ad dition, we generalize Sylvester Flows and propose Convolutional Sylvester Flows which are based on the generalization and the convolution exponential as basis c hange. Empirically, we show that the convolution exponential outperforms other l inear transformations in generative flows on CIFAR10 and the graph convolution e xponential improves the performance of graph normalizing flows. In addition, we show that Convolutional Sylvester Flows improve performance over residual flows as a generative flow model measured in log-likelihood.
**********

An Improved Analysis of Stochastic Gradient Descent with Momentum
Yanli Liu, Yuan Gao, Wotao Yin
SGD with momentum (SGDM) has been widely applied in many machine learning tasks,  and it is often applied with dynamic stepsizes and momentum weights tuned in a stagewise manner. Despite of its empirical advantage over SGD, the role of momen tum is still unclear in general since previous analyses on SGDM either provide w orse convergence bounds than those of SGD, or assume Lipschitz or quadratic obje ctives, which fail to hold in practice. Furthermore, the role of dynamic paramet ers has not been addressed. In this work, we show that SGDM converges as fast as  SGD for smooth objectives under both strongly convex and nonconvex settings. We  also prove that multistage strategy is beneficial for SGDM compared to using fi xed parameters. Finally, we verify these theoretical claims by numerical experim ents.
**********

Precise expressions for random projections: Low-rank approximation and randomize d Newton
Michal Derezinski, Feynman T. Liang, Zhenyu Liao, Michael W. Mahoney
It is often desirable to reduce the dimensionality of a large dataset
by projecting it onto a low-dimensional subspace.  Matrix sketching
has emerged as a powerful technique for performing such dimensionality
reduction very efficiently.  Even though there is an extensive
literature on the worst-case performance of sketching, existing
guarantees are typically very different from what is observed in
practice.  We exploit recent developments in the spectral analysis of
random matrices to develop novel techniques that provide provably
accurate expressions for the expected value of random projection
matrices obtained via sketching.  These expressions can be used to
characterize the performance of dimensionality reduction in a variety
of common machine learning tasks, ranging from low-rank approximation
to iterative stochastic optimization.  Our results apply to several
popular sketching methods, including Gaussian and Rademacher
sketches, and they enable precise analysis of these methods in terms of
spectral properties of the data.  Empirical results show that the
expressions we derive reflect the practical performance of these
sketching methods, down to lower-order effects and even constant factors.
**********

The MAGICAL Benchmark for Robust Imitation
Sam Toyer, Rohin Shah, Andrew Critch, Stuart Russell
Imitation Learning (IL) algorithms are typically evaluated in the same environme nt that was used to create demonstrations. This rewards precise reproduction of demonstrations in one particular environment, but provides little information ab out how robustly an algorithm can generalise the demonstrator's intent to substa

ntially different deployment settings. This paper presents the MAGICAL benchmark suite, which permits systematic evaluation of generalisation by quantifying robustness to different kinds of distribution shift that an IL algorithm is likely to encounter in practice. Using the MAGICAL suite, we confirm that existing IL algorithms overfit significantly to the context in which demonstrations are provided. We also show that standard methods for reducing overfitting are effective at creating narrow perceptual invariances, but are not sufficient to enable transfer to contexts that require substantially different behaviour, which suggests that new approaches will be needed in order to robustly generalise demonstrator intent. Code and data for the MAGICAL suite is available at https://github.com/qxcv/magical/
**********

X-CAL: Explicit Calibration for Survival Analysis
Mark Goldstein, Xintian Han, Aahlad Puli, Adler Perotte, Rajesh Ranganath
Survival analysis models the distribution of time until an event of interest, such as discharge from the hospital or admission to the ICU. When a model's predicted number of events within any time interval is similar to the observed number, it is called well-calibrated. A survival model's calibration can be measured using, for instance, distributional calibration (D-CALIBRATION) [Haider et al., 2020] which computes the squared difference between the observed and predicted number of events within different time intervals. Classically, calibration is addressed in post-training analysis. We develop explicit calibration (X-CAL), which turns D-CALIBRATION into a differentiable objective that can be used in survival modeling alongside maximum likelihood estimation and other objectives. X-CAL allows us to directly optimize calibration and strike a desired trade-off between predictive power and calibration. In our experiments, we fit a variety of shallow and deep models on simulated data, a survival dataset based on MNIST, on length-of-stay prediction using MIMIC-III data, and on brain cancer data from The Cancer Genome Atlas. We show that the models we study can be miscalibrated. We give experimental evidence on these datasets that X-CAL improves D-CALIBRATION without a large decrease in concordance or likelihood.
**********

Decentralized Accelerated Proximal Gradient Descent
Haishan Ye, Ziang Zhou, Luo Luo, Tong Zhang
Decentralized optimization has wide applications in machine learning, signal processing, and control.
    In this paper, we study the decentralized composite
    optimization problem with a non-smooth regularization term.
    Many  proximal gradient based decentralized algorithms have been proposed in the past.
    However, these algorithms do not achieve near optimal
    computational complexity and communication complexity.
    In this paper, we propose a new method which establishes the
    optimal computational
    complexity and a near optimal communication complexity.
    Our empirical study shows that the proposed algorithm outperforms existing state-of-the-art algorithms.
**********

Making Non-Stochastic Control (Almost) as Easy as Stochastic
Max Simchowitz
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

BERT Loses Patience: Fast and Robust Inference with Early Exit
Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, Furu Wei
In this paper, we propose Patience-based Early Exit, a straightforward yet effective inference method that can be used as a plug-and-play technique to simultaneously improve the efficiency and robustness of a pretrained language model (PLM)

. To achieve this, our approach couples an internal-classifier with each layer o
f a PLM and dynamically stops inference when the intermediate predictions of the
 internal classifiers do not change for a pre-defined number of steps. Our appro
ach improves inference efficiency as it allows the model to make a prediction wi
th fewer layers. Meanwhile, experimental results with an ALBERT model show that
 our method can improve the accuracy and robustness of the model by preventing it
 from overthinking and exploiting multiple classifiers for prediction, yielding
 a better accuracy-speed trade-off compared to existing early exit methods.
**********

Optimal and Practical Algorithms for Smooth and Strongly Convex Decentralized Op
timization
Dmitry Kovalev, Adil Salim, Peter Richtarik
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

BAIL: Best-Action Imitation Learning for Batch Deep Reinforcement Learning
Xinyue Chen, Zijian Zhou, Zheng Wang, Che Wang, Yanqiu Wu, Keith Ross
There has recently been a surge in research in batch Deep Reinforcement Learning
 (DRL), which aims for learning a high-performing policy from a given dataset wi
thout additional interactions with the environment. We propose a new algorithm,
Best-Action Imitation Learning (BAIL), which strives for both simplicity and per
formance. BAIL learns a V function, uses the V function to select actions it bel
ieves to be high-performing, and then uses those actions to train a policy netwo
rk using imitation learning. For the MuJoCo benchmark, we provide a comprehensiv
e experimental study of BAIL, comparing its performance to four other batch Q-le
arning and imitation-learning schemes for a large variety of batch datasets. Our
 experiments show that BAIL's performance is much higher than the other schemes,
 and is also computationally much faster than the batch Q-learning schemes.
**********

Regularizing Towards Permutation Invariance In Recurrent Models
Edo Cohen-Karlik, Avichai Ben David, Amir Globerson
In many machine learning problems the output should not depend on the order of t
he inputs. Such ``permutation invariant'' functions have been studied extensivel
y recently. Here we argue that temporal architectures such as RNNs are highly re
levant for such problems, despite the inherent dependence of RNNs on order. We s
how that RNNs can be regularized towards permutation invariance, and that this c
an result in compact models, as compared to non-recursive architectures.
Existing solutions (e.g., DeepSets) mostly suggest restricting the learning prob
lem to hypothesis classes which are permutation invariant by design. Our approac
h of enforcing permutation invariance via regularization gives rise to learning
functions which are "semi permutation invariant", e.g. invariant to some permuta
tions and not to others. Our approach relies on a novel form of stochastic regul
arization. We demonstrate that our method is beneficial compared to existing per
mutation invariant methods on synthetic and real world datasets.
**********

What Did You Think Would Happen? Explaining Agent Behaviour through Intended Out
comes
Herman Yau, Chris Russell, Simon Hadfield
We present a novel form of explanation for Reinforcement Learning, based around
the notion of intended outcome. These explanations describe the outcome an agent
 is trying to achieve by its actions. We provide a simple proof that general met
hods for post-hoc explanations of this nature are impossible in traditional rein
forcement learning. Rather, the information needed for the explanations must be
collected in conjunction with training the agent. We derive approaches designed
to extract local explanations based on intention for several variants of Q-funct
ion approximation and prove consistency between the explanations and the Q-value
s learned. We demonstrate our method on multiple reinforcement learning problems
, and provide code to help researchers introspecting their RL environments and a

lgorithms.
**********
Batch normalization provably avoids ranks collapse for randomly initialised deep networks
Hadi Daneshmand, Jonas Kohler, Francis Bach, Thomas Hofmann, Aurelien Lucchi
Randomly initialized neural networks are known to become harder to train with increasing depth, unless architectural enhancements like residual connections and batch normalization are used. We here investigate this phenomenon by revisiting the connection between random initialization in deep networks and spectral instabilities in products of random matrices. Given the rich literature on random matrices, it is not surprising to find that the rank of the intermediate representations in unnormalized networks collapses quickly with depth. In this work we highlight the fact that batch normalization is an effective strategy to avoid rank collapse for both linear and ReLU networks. Leveraging tools from Markov chain theory, we derive a meaningful lower rank bound in deep linear networks. Empirically, we also demonstrate that this rank robustness generalizes to ReLU nets. Finally, we conduct an extensive set of experiments on real-world data sets, which confirm that rank stability is indeed a crucial condition for training modern-day deep neural architectures.
**********
Choice Bandits
Arpit Agarwal, Nicholas Johnson, Shivani Agarwal
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********
What if Neural Networks had SVDs?
Alexander Mathiasen, Frederik Hvilshøj, Jakob Rødsgaard Jørgensen, Anshul Nasery, Davide Mottin
Various Neural Networks employ time-consuming matrix operations like matrix inversion.
Many such matrix operations are faster to compute given the Singular Value Decomposition (SVD).
Techniques from (Zhang et al., 2018; Mhammedi et al., 2017) allow using the SVD in Neural Networks without computing it.
In theory, the techniques can speed up matrix operations, however, in practice, they are not fast enough.
We present an algorithm that is fast enough to speed up several matrix operations.
The algorithm increases the degree of parallelism of an underlying matrix multiplication H*X where H is an orthogonal matrix represented by a product of Householder matrices.
**********
A Matrix Chernoff Bound for Markov Chains and Its Application to Co-occurrence Matrices
Jiezhong Qiu, Chi Wang, Ben Liao, Richard Peng, Jie Tang
We prove a Chernoff-type bound for sums of matrix-valued random variables sampled via a regular (aperiodic and irreducible) finite Markov chain. Specially, consider a random walk on a regular Markov chain and a Hermitian matrix-valued function on its state space. Our result gives exponentially decreasing bounds on the tail distributions of the extreme eigenvalues of the sample mean matrix. Our proof is based on the matrix expander (regular undirected graph) Chernoff bound [Garg et al. STOC '18] and scalar Chernoff-Hoeffding bounds for Markov chains [Chung et al. STACS '12].
**********
CoMIR: Contrastive Multimodal Image Representation for Registration
Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, Natasa Sladoje
We propose contrastive coding to learn shared, dense image representations, refe

rred to as CoMIRs (Contrastive Multimodal Image Representations). CoMIRs enable the registration of multimodal images where existing registration methods often fail due to a lack of sufficiently similar image structures. CoMIRs reduce the multimodal registration problem to a monomodal one, in which general intensity-based, as well as feature-based, registration algorithms can be applied. The method involves training one neural network per modality on aligned images, using a contrastive loss based on noise-contrastive estimation (InfoNCE). Unlike other contrastive coding methods, used for, e.g., classification, our approach generates image-like representations that contain the information shared between modalities. We introduce a novel, hyperparameter-free modification to InfoNCE, to enforce rotational equivariance of the learnt representations, a property essential to the registration task. We assess the extent of achieved rotational equivariance and the stability of the representations with respect to weight initialization, training set, and hyperparameter settings, on a remote sensing dataset of RGB and near-infrared images. We evaluate the learnt representations through registration of a biomedical dataset of bright-field and second-harmonic generation microscopy images; two modalities with very little apparent correlation. The proposed approach based on CoMIRs significantly outperforms registration of representations created by GAN-based image-to-image translation, as well as a state-of-the-art, application-specific method which takes additional knowledge about the data into account. Code is available at: https://github.com/MIDA-group/CoMIR.

**********

Ensuring Fairness Beyond the Training Data

Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, Daniel J. Hsu

We initiate the study of fair classifiers that are robust to perturbations in the training distribution. Despite recent progress, the literature on fairness has largely ignored the design of fair and robust classifiers. In this work, we develop classifiers that are fair not only with respect to the training distribution but also for a class of distributions that are weighted perturbations of the training samples. We formulate a min-max objective function whose goal is to minimize a distributionally robust training loss, and at the same time, find a classifier that is fair with respect to a class of distributions. We first reduce this problem to finding a fair classifier that is robust with respect to the class of distributions. Based on an online learning algorithm, we develop an iterative algorithm that provably converges to such a fair and robust solution.

Experiments on standard machine learning fairness datasets suggest that, compared to the state-of-the-art fair classifiers, our classifier retains fairness guarantees and test accuracy for a large class of perturbations on the test set. Furthermore, our experiments show that there is an inherent trade-off between fairness robustness and accuracy of such classifiers.

**********

How do fair decisions fare in long-term qualification?

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, Cheng Zhang

Although many fairness criteria have been proposed for decision making, their long-term impact on the well-being of a population remains unclear. In this work, we study the dynamics of population qualification and algorithmic decisions under a partially observed Markov decision problem setting. By characterizing the equilibrium of such dynamics, we analyze the long-term impact of static fairness constraints on the equality and improvement of group well-being. Our results show that static fairness constraints can either promote equality or exacerbate disparity depending on the driving factor of qualification transitions and the effect of sensitive attributes on feature distributions. We also consider possible interventions that can effectively improve group qualification or promote equality of group qualification. Our theoretical results and experiments on static real-world datasets with simulated dynamics show that our framework can be used to facilitate social science studies.

**********

Pre-training via Paraphrasing

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, Luke

Zettlemoyer
We introduce MARGE, a pre-trained sequence-to-sequence model learned with an unsupervised multi-lingual multi-document paraphrasing objective. MARGE provides an alternative to the dominant masked language modeling paradigm, where we self-supervise the \emph{reconstruction} of target text by \emph{retrieving} a set of related texts (in many languages) and conditioning on them to maximize the likelihood of generating the original. We show it is possible to jointly learn to do retrieval and reconstruction, given only a random initialization. The objective noisily captures aspects of paraphrase, translation, multi-document summarization, and information retrieval, allowing for strong zero-shot performance on several tasks. For example, with no additional task-specific training we achieve BLEU scores of up to 35.8 for document translation. We further show that fine-tuning gives strong performance on a range of discriminative and generative tasks in many languages, making MARGE the most generally applicable pre-training method to date.
**********
GCN meets GPU: Decoupling “When to Sample” from “How to Sample”
Morteza Ramezani, Weilin Cong, Mehrdad Mahdavi, Anand Sivasubramaniam, Mahmut Kandemir
Sampling-based methods promise scalability improvements when paired with stochastic gradient descent in training Graph Convolutional Networks (GCNs). While effective in alleviating the neighborhood explosion, due to bandwidth and memory bottlenecks, these methods lead to computational overheads in preprocessing and loading new samples in heterogeneous systems, which significantly deteriorate the sampling performance. By decoupling the frequency of sampling from the sampling strategy, we propose LazyGCN, a general yet effective framework that can be integrated with any sampling strategy to substantially improve the training time. The basic idea behind LazyGCN is to perform sampling periodically and effectively recycle the sampled nodes to mitigate data preparation overhead. We theoretically analyze the proposed algorithm and show that under a mild condition on the recycling size, by reducing the variance of inner layers, we are able to obtain the same convergence rate as the underlying sampling method. We also give corroborating empirical evidence on large real-world graphs, demonstrating that the proposed schema can significantly reduce the number of sampling steps and yield superior speedup without compromising the accuracy.
**********
Continual Learning of a Mixed Sequence of Similar and Dissimilar Tasks
Zixuan Ke, Bing Liu, Xingchang  Huang
Existing research on continual learning of a sequence of tasks focused on dealing with catastrophic forgetting, where the tasks are assumed to be dissimilar and have little shared knowledge. Some work has also been done to transfer previously learned knowledge to the new task when the tasks are similar and have shared knowledge. %However, in the most general case, a CL system not only should have the above two capabilities, but also the \textit{backward knowledge transfer} capability so that future tasks may help improve the past models whenever possible. To the best of our knowledge, no technique has been proposed to learn a sequence of mixed similar and dissimilar tasks that can deal with forgetting and also transfer knowledge forward and backward. This paper proposes such a technique to learn both types of tasks in the same network. For dissimilar tasks, the algorithm focuses on dealing with forgetting, and for similar tasks, the algorithm focuses on selectively transferring the knowledge learned from some similar previous tasks to improve the new task learning. Additionally, the algorithm automatically detects whether a new task is similar to any previous tasks. Empirical evaluation using sequences of mixed tasks demonstrates the effectiveness of the proposed model.
**********
All your loss are belong to Bayes
Christian Walder, Richard Nock
Loss functions are a cornerstone of machine learning and the starting point of most algorithms. Statistics and Bayesian decision theory have contributed, via pr

operness, to elicit over the past decades a wide set of admissible losses in sup
ervised learning, to which most popular choices belong (logistic, square, Matsus
hita, etc.). Rather than making a potentially biased ad hoc choice of the loss,
there has recently been a boost in efforts to fit the loss to the domain at hand
 while training the model itself. The key approaches fit a canonical link, a fun
ction which monotonically relates the closed unit interval to R and can provide
a proper loss via integration.
**********

HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks
Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, Kurt Ke
utzer
Quantization is an effective method for reducing memory footprint and inference
time of Neural Networks. However, ultra low precision quantization could lead to
 significant degradation in model accuracy. A promising method to address this i
s to perform mixed-precision quantization, where more sensitive layers are kept
at higher precision. However, the search space for a mixed-precision quantizatio
n is exponential in the number of layers. Recent work has proposed a novel Hessi
an based framework, with the aim of reducing this exponential search space by us
ing second-order information. While promising, this prior work has three major l
imitations: (i) they only use a heuristic metric based on top Hessian eigenvalue
 as a measure of sensitivity and do not consider the rest of the Hessian spectru
m; (ii) their approach only provides relative sensitivity of different layers an
d therefore requires a manual selection of the mixed-precision setting; and (iii
) they do not consider mixed-precision activation quantization. Here, we present
 HAWQ-V2 which addresses these shortcomings. For (i), we theoretically prove tha
t the right sensitivity metric is the average Hessian trace, instead of just top
 Hessian eigenvalue. For (ii), we develop a Pareto frontier based method for aut
omatic bit precision selection of different layers without any manual interventi
on. For (iii), we develop the first Hessian based analysis for mixed-precision a
ctivation quantization, which is very beneficial for object detection. We show t
hat HAWQ-V2 achieves new state-of-the-art results for a wide range of tasks. In
particular, we present quantization results for InceptionV3, ResNet50, and Squee
zeNext, all without any manual bit selection. Furthermore, we present results fo
r object detection on Microsoft COCO, where we achieve 2.6 higher mAP than direc
t uniform quantization and 1.6 higher mAP than the recently proposed method of F
QN, with a smaller model size of 17.9MB.
**********

Sample-Efficient Reinforcement Learning of Undercomplete POMDPs
Chi Jin, Sham Kakade, Akshay Krishnamurthy, Qinghua 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Non-Convex SGD Learns Halfspaces with Adversarial Label Noise
Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, Nikos Zarifis
Requests for name changes in the electronic proceedings will be accepted with 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 Tight Lower Bound and Efficient Reduction for Swap Regret
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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

DisCor: Corrective Feedback in Reinforcement Learning via Distribution Correctio
n

Aviral Kumar, Abhishek Gupta, Sergey Levine
Deep reinforcement learning can learn effective policies for a wide range of tas
ks, but is notoriously difficult to use due to instability and sensitivity to hy
perparameters. The reasons for this remain unclear. In this paper, we study how
RL methods based on bootstrapping-based Q-learning can suffer from a pathologica
l interaction between function approximation and the data distribution used to t
rain the Q-function: with standard supervised learning, online data collection s
hould induce corrective feedback, where new data corrects mistakes in old predic
tions. With dynamic programming methods like Q-learning, such feedback may be ab
sent. This can lead to potential instability, sub-optimal convergence, and poor
results when learning from noisy, sparse or delayed rewards. Based on these obse
rvations, we propose a new algorithm, DisCor, which explicitly optimizes for dat
a distributions that can correct for accumulated errors in the value function. D
isCor computes a tractable approximation to the distribution that optimally indu
ces corrective feedback, which we show results in reweighting samples based on t
he estimated accuracy of their target values. Using this distribution for traini
ng, DisCor results in substantial improvements in a range of challenging RL sett
ings, such as multi-task learning and learning from noisy reward signals.
**********

OTLDA: A Geometry-aware Optimal Transport Approach for Topic Modeling
Viet Huynh, He Zhao, Dinh Phung
We present an optimal transport framework for learning topics from textual data.
 While the celebrated Latent Dirichlet allocation (LDA) topic model and its vari
ants have been applied to many disciplines, they mainly focus on word-occurrence
s and neglect to incorporate semantic regularities in language. Even though rece
nt works have tried to exploit the semantic relationship between words to bridge
 this gap, however, these models which are usually extensions of LDA or Dirichle
t Multinomial mixture (DMM) are tailored to deal effectively with either regular
 or short documents. The optimal transport distance provides an appealing tool t
o incorporate the geometry of word semantics into it. Moreover, recent developme
nts on efficient computation of optimal transport distance also promote its appl
ication in topic modeling. In this paper we ground on optimal transport theory t
o naturally exploit the geometric structures of semantically related words in em
bedding spaces which leads to more interpretable learned topics. Comprehensive e
xperiments illustrate that the proposed framework outperforms competitive approa
ches in terms of topic coherence on assorted text corpora which include both lon
g and short documents. The representation of learned topic also leads to better
accuracy on classification downstream tasks, which is considered as an extrinsic
 evaluation.
**********

Measuring Robustness to Natural Distribution Shifts in Image Classification
Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, Lud
wig Schmidt
We study how robust current ImageNet models are to distribution shifts arising f
rom natural variations in datasets. Most research on robustness focuses on synth
etic image perturbations (noise, simulated weather artifacts, adversarial exampl
es, etc.), which leaves open how robustness on synthetic distribution shift rela
tes to distribution shift arising in real data. Informed by an evaluation of 204
 ImageNet models in 213 different test conditions, we find that there is often l
ittle to no transfer of robustness from current synthetic to natural distributio
n shift. Moreover, most current techniques provide no robustness to the natural
distribution shifts in our testbed. The main exception is training on larger and
 more diverse datasets, which in multiple cases increases robustness, but is sti
ll far from closing the performance gaps. Our results indicate that distribution
 shifts arising in real data are currently an open research problem.
**********

Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayes
ian Inference
Disi Ji, Padhraic Smyth, Mark Steyvers
Group fairness is measured via parity of quantitative metrics across different p

rotected demographic groups. In this paper, we investigate the problem of reliab ly assessing group fairness metrics when labeled examples are few but unlabeled examples are plentiful. We propose a general Bayesian framework that can augment labeled data with unlabeled data to produce more accurate and lower-variance es timates compared to methods based on labeled data alone. Our approach estimates calibrated scores (for unlabeled examples) of each group using a hierarchical la tent variable model conditioned on labeled examples. This in turn allows for inf erence of posterior distributions for an array of group fairness metrics with a notion of uncertainty. We demonstrate that our approach leads to significant and consistent reductions in estimation error across multiple well-known fairness d atasets, sensitive attributes, and predictive models. The results clearly show t he benefits of using both unlabeled data and Bayesian inference in assessing whe ther a prediction model is fair or not.
**********

RandAugment: Practical Automated Data Augmentation with a Reduced Search Space
Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, Quoc Le
Recent work on automated data augmentation strategies has led to state-of-the-ar t results in image classification and object detection. An obstacle to a large-s cale adoption of these methods is that they require a separate and expensive sea rch phase. A common way to overcome the expense of the search phase was to use a smaller proxy task. However, it was not clear if the optimized hyperparameters found on the proxy task are also optimal for the actual task. In this work, we r ethink the process of designing automated data augmentation strategies. We find that while previous work required searching for many augmentation parameters (e. g. magnitude and probability) independently for each augmentation operation, it is sufficient to only search for a single parameter that jointly controls all op erations. Hence, we propose a search space that is vastly smaller (e.g. from 10^ 32 to 10^2 potential candidates). The smaller search space significantly reduces the computational expense of automated data augmentation and permits the remova l of a separate proxy task. Despite the simplifications, our method achieves sta te-of-the-art performance on CIFAR-10, SVHN, and ImageNet. On EfficientNet-B7, w e achieve 84.7% accuracy, a 1.0% increase over baseline augmentation and a 0.4% improvement over AutoAugment on the ImageNet dataset. On object detection, the s ame method used for classification leads to 1.0-1.3% improvement over the baseli ne augmentation method on COCO. Code is available online.
**********

Asymptotic normality and confidence intervals for derivatives of 2-layers neural network in the random features model
Yiwei Shen, Pierre C Bellec
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

DisARM: An Antithetic Gradient Estimator for Binary Latent Variables
Zhe Dong, Andriy Mnih, George Tucker
Training models with discrete latent variables is challenging due to the difficu lty of estimating the gradients accurately. Much of the recent progress has been achieved by taking advantage of continuous relaxations of the system, which are not always available or even possible. The Augment-REINFORCE-Merge (ARM) estima tor provides an alternative that, instead of relaxation, uses continuous augment ation. Applying antithetic sampling over the augmenting variables yields a relat ively low-variance and unbiased estimator applicable to any model with binary la tent variables. However, while antithetic sampling reduces variance, the augment ation process increases variance. We show that ARM can be improved by analytical ly integrating out the randomness introduced by the augmentation process, guaran teeing substantial variance reduction. Our estimator, DisARM, is simple to imple ment and has the same computational cost as ARM. We evaluate DisARM on several g enerative modeling benchmarks and show that it consistently outperforms ARM and a strong independent sample baseline in terms of both variance and log-likelihoo

d. Furthermore, we propose a local version of DisARM designed for optimizing the multi-sample variational bound, and show that it outperforms VIMCO, the current state-of-the-art method.
**********

Variational Inference for Graph Convolutional Networks in the Absence of Graph Data and Adversarial Settings
Pantelis Elinas, Edwin V. Bonilla, Louis Tiao
We propose a framework that lifts the capabilities of graph convolutional networks (GCNs) to scenarios where no input graph is given and increases their robustness to adversarial attacks. We formulate a joint probabilistic model that considers a prior distribution over graphs along with a GCN-based likelihood and develop a stochastic variational inference algorithm to estimate the graph posterior and the GCN parameters jointly. To address the problem of propagating gradients through latent variables drawn from discrete distributions, we use their continuous relaxations known as Concrete distributions. We show that, on real datasets, our approach can outperform state-of-the-art Bayesian and non-Bayesian graph neural network algorithms on the task of semi-supervised classification in the absence of graph data and when the network structure is subjected to adversarial perturbations.
**********

Supervised Contrastive Learning
Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan
Contrastive learning applied to self-supervised representation learning has seen a resurgence in recent years, leading to state of the art performance in the unsupervised training of deep image models. Modern batch contrastive approaches subsume or significantly outperform traditional contrastive losses such as triplet, max-margin and the N-pairs loss. In this work, we extend the self-supervised batch contrastive approach to the fully-supervised setting, allowing us to effectively leverage label information. Clusters of points belonging to the same class are pulled together in embedding space, while simultaneously pushing apart clusters of samples from different classes. We analyze two possible versions of the supervised contrastive (SupCon) loss, identifying the best-performing formulation of the loss. On ResNet-200, we achieve top-1 accuracy of 81.4% on the ImageNet dataset, which is 0.8% above the best number reported for this architecture. We show consistent outperformance over cross-entropy on other datasets and two ResNet variants. The loss shows benefits for robustness to natural corruptions, and is more stable to hyperparameter settings such as optimizers and data augmentations. In reduced data settings, it outperforms cross-entropy significantly. Our loss function is simple to implement and reference TensorFlow code is released at https://t.ly/supcon.
**********

Learning Optimal Representations with the Decodable Information Bottleneck
Yann Dubois, Douwe Kiela, David J. Schwab, Ramakrishna Vedantam
We address the question of characterizing and finding optimal representations for supervised learning. Traditionally, this question has been tackled using the Information Bottleneck, which compresses the inputs while retaining information about the targets, in a decoder-agnostic fashion. In machine learning, however, our goal is not compression but rather generalization, which is intimately linked to the predictive family or decoder of interest (e.g. linear classifier). We propose the Decodable Information Bottleneck (DIB) that considers information retention and compression from the perspective of the desired predictive family. As a result, DIB gives rise to representations that are optimal in terms of expected test performance and can be estimated with guarantees. Empirically, we show that the framework can be used to enforce a small generalization gap on downstream classifiers and to predict the generalization ability of neural networks.
**********

Meta-trained agents implement Bayes-optimal agents
Vladimir Mikulik, Grégoire Delétang, Tom McGrath, Tim Genewein, Miljan Martic, Shane Legg, Pedro Ortega

Memory-based meta-learning is a powerful technique to build agents that adapt fa st to any task within a target distribution. A previous theoretical study has ar gued that this remarkable performance is because the meta-training protocol ince ntivises agents to behave Bayes-optimally. We empirically investigate this claim on a number of prediction and bandit tasks. Inspired by ideas from theoretical computer science, we show that meta-learned and Bayes-optimal agents not only be have alike, but they even share a similar computational structure, in the sense that one agent system can approximately simulate the other. Furthermore, we show that Bayes-optimal agents are fixed points of the meta-learning dynamics. Our r esults suggest that memory-based meta-learning is a general technique for numeri cally approximating Bayes-optimal agents; that is, even for task distributions f or which we currently don't possess tractable models.

**********

Learning Agent Representations for Ice Hockey

Guiliang Liu, Oliver Schulte, Pascal Poupart, Mike Rudd, Mehrsan Javan

Team sports is a new application domain for agent modeling with high real-world impact. A fundamental challenge for modeling professional players is their large number (over 1K), which includes many bench players with sparse participation i n a game season. The diversity and sparsity of player observations make it diffi cult to extend previous agent representation models to the sports domain. This p aper develops a new approach for agent representations, based on a Markov game m odel, that is tailored towards applications in professional ice hockey. We intro duce a novel player representation via player generation framework where a varia tional encoder embeds player information with latent variables. The encoder lear ns a context-specific shared prior to induce a shrinkage effect for the posterio r player representations, allowing it to share statistical information across pl ayers with different participations. To model the play dynamics in sequential sp orts data, we design a Variational Recurrent Ladder Agent Encoder (VaRLAE). It learns a contextualized player representation with a hierarchy of latent variabl es that effectively prevents latent posterior collapse. We validate our player r epresentations in major sports analytics tasks. Our experimental results, based on a large dataset that contains over 4.5M events, show state-of-the-art perform ance for our VarLAE on facilitating 1) identifying the acting player, 2) estimat ing expected goals, and 3) predicting the final score difference.

**********

Weak Form Generalized Hamiltonian Learning

Kevin Course, Trefor Evans, Prasanth Nair

We present a method for learning generalized Hamiltonian decompositions of ordin ary differential equations given a set of noisy time series measurements. Our me thod simultaneously learns a continuous time model and a scalar energy function for a general dynamical system. Learning predictive models in this form allows o ne to place strong, high-level, physics inspired priors onto the form of the lea rnt governing equations for general dynamical systems. Moreover, having shown ho w our method extends and unifies some previous work in deep learning with physic s inspired priors, we present a novel method for learning continuous time models from the weak form of the governing equations which is less computationally tax ing than standard adjoint methods.

**********

Neural Non-Rigid Tracking

Aljaz Bozic, Pablo Palafox, Michael Zollhöfer, Angela Dai, Justus Thies, Matthi as Niessner

We introduce a novel, end-to-end learnable, differentiable non-rigid tracker tha t enables state-of-the-art non-rigid reconstruction by a learned robust optimiza tion. Given two input RGB-D frames of a non-rigidly moving object, we employ a c onvolutional neural network to predict dense correspondences and their confidenc es. These correspondences are used as constraints in an as-rigid-as-possible (AR AP) optimization problem. By enabling gradient back-propagation through the weig hted non-linear least squares solver, we are able to learn correspondences and confidences in an end-to-end manner such that they are optimal for the task of n on-rigid tracking. Under this formulation, correspondence confidences can be lea

rned via self-supervision, informing a learned robust optimization, where outlie rs and wrong correspondences are automatically down-weighted to enable effective tracking. Compared to state-of-the-art approaches, our algorithm shows improved reconstruction performance, while simultaneously achieving 85 times faster corr espondence prediction than comparable deep-learning based methods.
**********

Collegial Ensembles
Etai Littwin, Ben Myara, Sima Sabah, Joshua Susskind, Shuangfei Zhai, Oren Golan
Modern neural network performance typically improves as model size increases. A recent line of research on the Neural Tangent Kernel (NTK) of over-parameterized networks indicates that the improvement with size increase is a product of a be tter conditioned loss landscape. In this work, we investigate a form of over-par ameterization achieved through ensembling, where we define collegial ensembles ( CE) as the aggregation of multiple independent models with identical architectur es, trained as a single model. We show that the optimization dynamics of CE simp lify dramatically when the number of models in the ensemble is large, resembling the dynamics of wide models, yet scale much more favorably. We use recent theor etical results on the finite width corrections of the NTK to perform efficient a rchitecture search in a space of finite width CE that aims to either minimize ca pacity, or maximize trainability under a set of constraints. The resulting ensem bles can be efficiently implemented in practical architectures using group convo lutions and block diagonal layers. Finally, we show how our framework can be use d to analytically derive optimal group convolution modules originally found usin g expensive grid searches, without having to train a single model.
**********

ICNet: Intra-saliency Correlation Network for Co-Saliency Detection
Wen-Da Jin, Jun Xu, Ming-Ming Cheng, Yi Zhang, Wei Guo
Intra-saliency and inter-saliency cues have been extensively studied for co-sali ency detection (Co-SOD). Model-based methods produce coarse Co-SOD results due t o hand-crafted intra- and inter-saliency features. Current data-driven models ex ploit inter-saliency cues, but undervalue the potential power of intra-saliency cues. In this paper, we propose an Intra-saliency Correlation Network (ICNet) to extract intra-saliency cues from the single image saliency maps (SISMs) predict ed by any off-the-shelf SOD method, and obtain inter-saliency cues by correlatio n techniques. Specifically, we adopt normalized masked average pooling (NMAP) to extract latent intra-saliency categories from the SISMs and semantic features a s intra cues. Then we employ a correlation fusion module (CFM) to obtain inter c ues by exploiting correlations between the intra cues and single-image features. To improve Co-SOD performance, we propose a category-independent rearranged sel f-correlation feature (RSCF) strategy. Experiments on three benchmarks show that our ICNet outperforms previous state-of-the-art methods on Co-SOD. Ablation stu dies validate the effectiveness of our contributions. The PyTorch code is availa ble at https://github.com/blanclist/ICNet.
**********

Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows
Cheng Zhang
Variational Bayesian phylogenetic inference (VBPI) provides a promising general variational framework for efficient estimation of phylogenetic posteriors. Howev er, the current diagonal Lognormal branch length approximation would significant ly restrict the quality of the approximating distributions. In this paper, we pr opose a new type of VBPI, VBPI-NF, as a first step to empower phylogenetic poste rior estimation with deep learning techniques. By handling the non-Euclidean bra nch length space of phylogenetic models with carefully designed permutation equi variant transformations, VBPI-NF uses normalizing flows to provide a rich family of flexible branch length distributions that generalize across different tree t opologies. We show that VBPI-NF significantly improves upon the vanilla VBPI on a benchmark of challenging real data Bayesian phylogenetic inference problems. F urther investigation also reveals that the structured parameterization in those permutation equivariant transformations can provide additional amortization bene fit.

**********
Deep Metric Learning with Spherical Embedding
Dingyi Zhang, Yingming Li, Zhongfei Zhang
Deep metric learning has attracted much attention in recent years, due to seamle ssly combining the distance metric learning and deep neural network. Many endeav ors are devoted to design different pair-based angular loss functions, which de couple the magnitude and direction information for embedding vectors and ensure the training and testing measure consistency. However, these traditional angular losses cannot guarantee that all the sample embeddings are on the surface of th e same hypersphere during the training stage, which would result in unstable gra dient in batch optimization and may influence the quick convergence of the embed ding learning. In this paper, we first investigate the effect of the embedding n orm for deep metric learning with angular distance, and then propose a spherical embedding constraint (SEC) to regularize the distribution of the norms. SEC ada ptively adjusts the embeddings to fall on the same hypersphere and performs more balanced direction update. Extensive experiments on deep metric learning, face recognition, and contrastive self-supervised learning show that the SEC-based an gular space learning strategy significantly improves the performance of the stat e-of-the-art.
**********
Preference-based Reinforcement Learning with Finite-Time Guarantees
Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, Artur Dubrawski
Requests for name changes in the electronic proceedings will be accepted with 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.
**********
AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients
Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C. Tatikonda, Nicha Dvornek, Xeno phon Papademetris, James Duncan
Most popular optimizers for deep learning can be broadly categorized as adaptive methods (e.g.~Adam) and accelerated schemes (e.g.~stochastic gradient descent ( SGD) with momentum).
For many models such as convolutional neural networks (CNNs), adaptive methods t ypically converge faster but generalize worse compared to SGD; for complex setti ngs such as generative adversarial networks (GANs), adaptive methods are typical ly the default because of their stability. We propose AdaBelief to simultaneousl y achieve three goals: fast convergence as in adaptive methods, good generalizat ion as in SGD, and training stability. The intuition for AdaBelief is to adapt t he stepsize according to the "belief" in the current gradient direction.
Viewing the exponential moving average (EMA) of the noisy gradient as the predic tion of the gradient at the next time step, if the observed gradient greatly dev iates from the prediction, we distrust the current observation and take a small step; if the observed gradient is close to the prediction, we trust it and take a large step.
We validate AdaBelief in extensive experiments, showing that it outperforms othe r methods with fast convergence and high accuracy on image classification and la nguage modeling. Specifically, on ImageNet, AdaBelief achieves comparable accura cy to SGD. Furthermore, in the training of a GAN on Cifar10, AdaBelief demonstra tes high stability and improves the quality of generated samples compared to a w ell-tuned Adam optimizer.
**********
Interpretable Sequence Learning for Covid-19 Forecasting
Sercan Arik, Chun-Liang Li, Jinsung Yoon, Rajarishi Sinha, Arkady Epshteyn, Long Le, Vikas Menon, Shashank Singh, Leyou Zhang, Martin Nikoltchev, Yash Sonthali a, Hootan Nakhost, Elli Kanal, Tomas Pfister
We propose a novel approach that integrates machine learning into compartmental disease modeling (e.g., SEIR) to predict the progression of COVID-19. Our model is explainable by design as it explicitly shows how different compartments evolv e and it uses interpretable encoders to incorporate covariates and improve perfo

rmance. Explainability is valuable to ensure that the model's forecasts are cred ible to epidemiologists and to instill confidence in end-users such as policy ma kers and healthcare institutions.  Our model can be applied at different geograp hic resolutions, and we demonstrate it for states and counties in the United Sta tes. We show that our model provides more accurate forecasts compared to the alt ernatives, and that it provides qualitatively meaningful explanatory insights.
**********

Off-policy Policy Evaluation For Sequential Decisions Under Unobserved Confoundi ng

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, Emma Brunskill

When observed decisions depend only on observed features, off-policy policy eval uation (OPE) methods for sequential decision problems can estimate the performan ce of evaluation policies before deploying them. However, this assumption is fre quently violated due to unobserved confounders, unrecorded variables that impact  both the decisions and their outcomes. We assess robustness of OPE methods unde r unobserved confounding by developing worst-case bounds on the performance of a n evaluation policy. When unobserved confounders can affect every decision in an  episode, we demonstrate that even small amounts of per-decision confounding can  heavily bias OPE methods. Fortunately, in a number of important settings found in healthcare, policy-making, and technology, unobserved confounders may directl y affect only one of the many decisions made, and influence future decisions/rew ards only through the directly affected decision. Under this less pessimistic mo del of one-decision confounding, we propose an efficient loss-minimization-based  procedure for computing worst-case bounds, and prove its statistical consistenc y. On simulated healthcare examples---management of sepsis and interventions for  autistic children---where this is a reasonable model, we demonstrate that our m ethod invalidates non-robust results and provides meaningful certificates of rob ustness, allowing reliable selection of policies under unobserved confounding.
**********

Modern Hopfield Networks and Attention for Immune Repertoire Classification

Michael Widrich, Bernhard Schäfl, Milena Pavlovi■, Hubert Ramsauer, Lukas Gruber , Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, Günter Klambauer

A central mechanism in machine learning is to identify, store, and recognize pat terns. How to learn, access, and retrieve such patterns is crucial in Hopfield n etworks and the more recent transformer architectures. We show that the attentio n mechanism of transformer architectures is actually the update rule of modern H opfield networks that can store exponentially many patterns. We exploit this hig h storage capacity of modern Hopfield networks to solve a challenging multiple i nstance learning (MIL) problem in computational biology: immune repertoire class ification. In immune repertoire classification, a vast number of immune receptor s are used to predict the immune status of an individual. This constitutes a MIL  problem with an unprecedentedly massive number of instances, two orders of magn itude larger than currently considered problems, and with an extremely low witne ss rate. Accurate and interpretable machine learning methods solving this proble m could pave the way towards new vaccines and therapies, which is currently a ve ry relevant research topic intensified by the COVID-19 crisis. In this work, we present our novel method DeepRC that integrates transformer-like attention, or e quivalently modern Hopfield networks, into deep learning architectures for massi ve MIL such as immune repertoire classification. We demonstrate that DeepRC outp erforms all other methods with respect to predictive performance on large-scale experiments including simulated and real-world virus infection data and enables the extraction of sequence motifs that are connected to a given disease class. S ource code and datasets: https://github.com/ml-jku/DeepRC
**********

One Ring to Rule Them All: Certifiably Robust Geometric Perception with Outliers

Heng Yang, Luca Carlone

We propose the first general and practical framework to design certifiable algor ithms for robust geometric perception in the presence of a large amount of outli ers. We investigate the use of a truncated least squares (TLS) cost function, wh

ich is known to be robust to outliers, but leads to hard, nonconvex, and nonsmoo th optimization problems. Our first contribution is to show that –for a broad cl ass of geometric perception problems– TLS estimation can be reformulated as an o ptimization over the ring of polynomials and Lasserre's hierarchy of convex mome nt relaxations is empirically tight at the minimum relaxation order (i.e., certi fiably obtains the global minimum of the nonconvex TLS problem). Our second cont ribution is to exploit the structural sparsity of the objective and constraint p olynomials and leverage basis reduction to significantly reduce the size of the semidefinite program (SDP) resulting from the moment relaxation, without comprom ising its tightness. Our third contribution is to develop scalable dual optimali ty certifiers from the lens of sums-of-squares (SOS) relaxation, that can comput e the suboptimality gap and possibly certify global optimality of any candidate solution (e.g., returned by fast heuristics such as RANSAC or graduated non-conv exity). Our dual certifiers leverage Douglas-Rachford Splitting to solve a conve x feasibility SDP. Numerical experiments across different perception problems, i ncluding single rotation averaging, shape alignment, 3D point cloud and mesh reg istration, and high-integrity satellite pose estimation, demonstrate the tightne ss of our relaxations, the correctness of the certification, and the scalability of the proposed dual certifiers to large problems, beyond the reach of current SDP solvers.

**********

Task-Robust Model-Agnostic Meta-Learning

Liam Collins, Aryan Mokhtari, 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-auth ors prior to requesting a name change in the electronic proceedings.

**********

R-learning in actor-critic model offers a biologically relevant mechanism for se quential decision-making

Sergey Shuvaev, Sarah Starosta, Duda Kvitsiani, Adam Kepecs, Alexei Koulakov

In real-world settings, we repeatedly decide whether to pursue better conditions or to keep things unchanged. Examples include time investment, employment, ente rtainment preferences etc. How do we make such decisions? To address this questi on, the field of behavioral ecology has developed foraging paradigms – the model settings in which human and non-human subjects decided when to leave depleting food resources. Foraging theory, represented by the marginal value theorem (MVT) , provided accurate average-case stay-or-leave rules consistent with behaviors o f subjects towards depleting resources. Yet, the algorithms underlying individua l choices and ways to learn such algorithms remained unclear. In this work, we b uild interpretable deep actor-critic models to show that R-learning – a reinforc ement learning (RL) approach balancing short-term and long-term rewards – is con sistent with the way real-life agents may learn making stay-or-leave decisions. Specifically we show that deep R-learning predicts choice patterns consistent wi th behavior of mice in foraging tasks; its TD error, the training signal in our model, correlates with dopamine activity of ventral tegmental area (VTA) neurons in the brain. Our theoretical and experimental results show that deep R-learnin g agents leave depleting reward resources when reward intake rates fall below th eir exponential averages over past trials. This individual-case decision rule, l earned within RL and matching the MVT on average, bridges the gap between these major approaches to sequential decision-making. We further argue that our propos ed decision rule, resulting from R-learning and consistent with animals' behavio r, is Bayes optimal in dynamic real-world environments. Overall, our work links available sequential decision-making theories including the MVT, RL, and Bayesia n approaches to propose the learning mechanism and an optimal decision rule for sequential stay-or-leave choices in natural environments.

**********

Revisiting Frank-Wolfe for Polytopes: Strict Complementarity and Sparsity

Dan Garber

In recent years it was proved that simple modifications of the classical Frank-W

olfe algorithm (aka conditional gradient algorithm) for smooth convex minimizati on over convex and compact polytopes, converge with linear rate, assuming the ob jective function has the quadratic growth property. However, the rate of these m ethods depends explicitly on the dimension of the problem which cannot explain t heir empirical success for large scale problems. In this paper we first demonstr ate that already for very simple problems and even when the optimal solution lie s on a low-dimensional face of the polytope, such dependence on the dimension ca nnot be avoided in worst case. We then revisit the addition of a strict compleme ntarity assumption already considered in Wolfe's classical book \cite{Wolfe1970} , and prove that under this condition, the Frank-Wolfe method with away-steps an d line-search converges linearly with rate that depends explicitly only on the d imension of the optimal face, hence providing a significant improvement in case the optimal solution is sparse. We motivate this strict complementarity conditio n by proving that it implies sparsity-robustness of optimal solutions to noise.
**********

Fast Convergence of Langevin Dynamics on Manifold: Geodesics meet Log-Sobolev
Xiao Wang, Qi Lei, Ioannis Panageas
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Tensor Completion Made Practical
Allen Liu, Ankur Moitra
Tensor completion is a natural higher-order generalization of matrix completion where the goal is to recover a low-rank tensor from sparse observations of its e ntries. Existing algorithms are either heuristic without provable guarantees, ba sed on solving large semidefinite programs which are impractical to run, or make strong assumptions such as requiring the factors to be nearly orthogonal. In t his paper we introduce a new variant of alternating minimization, which in turn is inspired by understanding how the progress measures that guide convergence of alternating minimization in the matrix setting need to be adapted to the tensor setting. We show strong provable guarantees, including showing that our algorit hm converges linearly to the true tensors even when the factors are highly corre lated and can be implemented in nearly linear time. Moreover our algorithm is al so highly practical and we show that we can complete third order tensors with a thousand dimensions from observing a tiny fraction of its entries. In contrast, and somewhat surprisingly, we show that the standard version of alternating mini mization, without our new twist, can converge at a drastically slower rate in pr actice.
**********

Optimization and Generalization Analysis of Transduction through Gradient Boosti ng and Application to Multi-scale Graph Neural Networks
Kenta Oono, Taiji Suzuki
It is known that the current graph neural networks (GNNs) are difficult to make themselves deep due to the problem known as over-smoothing. Multi-scale GNNs are a promising approach for mitigating the over-smoothing problem. However, there is little explanation of why it works empirically from the viewpoint of learning theory. In this study, we derive the optimization and generalization guarantees of transductive learning algorithms that include multi-scale GNNs. Using the bo osting theory, we prove the convergence of the training error under weak learnin g-type conditions. By combining it with generalization gap bounds in terms of tr ansductive Rademacher complexity, we show that a test error bound of a specific type of multi-scale GNNs that decreases corresponding to the number of node aggr egations under some conditions. Our results offer theoretical explanations for t he effectiveness of the multi-scale structure against the over-smoothing problem . We apply boosting algorithms to the training of multi-scale GNNs for real-worl d node prediction tasks. We confirm that its performance is comparable to existi ng GNNs, and the practical behaviors are consistent with theoretical observation s. Code is available at https://github.com/delta2323/GB-GNN.

**********
Content Provider Dynamics and Coordination in Recommendation Ecosystems
Omer Ben-Porat, Itay Rosenberg, Moshe Tennenholtz
Recommendation Systems like YouTube are vibrant ecosystems with two types of users: Content consumers (those who watch videos) and content providers (those who create videos). While the computational task of recommending relevant content is largely solved, designing a system that guarantees high social welfare for \textit{all} stakeholders is still in its infancy. In this work, we investigate the dynamics of content creation using a game-theoretic lens. Employing a stylized model that was recently suggested by other works, we show that the dynamics will always converge to a pure Nash Equilibrium (PNE), but the convergence rate can be exponential. We complement the analysis by proposing an efficient PNE computation algorithm via a combinatorial optimization problem that is of independent interest.
**********
Almost Surely Stable Deep Dynamics
Nathan Lawrence, Philip Loewen, Michael Forbes, Johan Backstrom, Bhushan Gopaluni
We introduce a method for learning provably stable deep neural network based dynamic models from observed data. Specifically, we consider discrete-time stochastic dynamic models, as they are of particular interest in practical applications such as estimation and control. However, these aspects exacerbate the challenge of guaranteeing stability. Our method works by embedding a Lyapunov neural network into the dynamic model, thereby inherently satisfying the stability criterion. To this end, we propose two approaches and apply them in both the deterministic and stochastic settings: one exploits convexity of the Lyapunov function, while the other enforces stability through an implicit output layer. We demonstrate the utility of each approach through numerical examples.
**********
Experimental design for MRI by greedy policy search
Tim Bakker, Herke van Hoof, Max Welling
In today’s clinical practice, magnetic resonance imaging (MRI) is routinely accelerated through subsampling of the associated Fourier domain. Currently, the construction of these subsampling strategies - known as experimental design - relies primarily on heuristics. We propose to learn experimental design strategies for accelerated MRI with policy gradient methods. Unexpectedly, our experiments show that a simple greedy approximation of the objective leads to solutions nearly on-par with the more general non-greedy approach. We offer a partial explanation for this phenomenon rooted in greater variance in the non-greedy objective's gradient estimates, and experimentally verify that this variance hampers non-greedy models in adapting their policies to individual MR images. We empirically show that this adaptivity is key to improving subsampling designs.
**********
Expert-Supervised Reinforcement Learning for Offline Policy Learning and Evaluation
Aaron Sonabend, Junwei Lu, Leo Anthony Celi, Tianxi Cai, Peter Szolovits
Offline Reinforcement Learning (RL) is a promising approach for learning optimal policies in environments where direct exploration is expensive or unfeasible. However, the adoption of such policies in practice is often challenging, as they are hard
to interpret within the application context, and lack measures of uncertainty for the
learned policy value and its decisions. To overcome these issues, we propose an Expert-Supervised RL (ESRL) framework which uses uncertainty quantification
for offline policy learning. In particular, we have three contributions: 1) the method
can learn safe and optimal policies through hypothesis testing, 2) ESRL allows for
different levels of risk averse implementations tailored to the application context,

and finally, 3) we propose a way to interpret ESRL’s policy at every state throu
gh
posterior distributions, and use this framework to compute off-policy value func
tion
posteriors. We provide theoretical guarantees for our estimators and regret boun
ds
consistent with Posterior Sampling for RL (PSRL). Sample efficiency of ESRL
is independent of the chosen risk aversion threshold and quality of the behavior
policy.
**********

ColdGANs: Taming Language GANs with Cautious Sampling Strategies
Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo
Staiano
Training regimes based on Maximum Likelihood Estimation (MLE) suffer from known
limitations, often leading to poorly generated text sequences that lack of coher
ence, factualness, and are prone to repetitions. At the root of these limitation
s is the mismatch between training and inference, i.e. the so-called exposure bi
as. Another problem lies in considering only the reference text as correct, whil
e in practice several alternative formulations could be as good.
**********

Hedging in games: Faster convergence of external and swap regrets
Xi Chen, Binghui Peng
We consider the setting where players run the Hedge algorithm or its optimistic
variant \cite{syrgkanis2015fast} to play an n-action game repeatedly for T round
s.
1) For two-player games, we show that the regret of optimistic Hedge decays at \
tilde{O}( 1/T ^{5/6} ), improving the previous bound O(1/T^{3/4}) by \cite{syrgk
anis2015fast}.
2) In contrast, we show that the convergence rate of vanilla Hedge is no better
than \tilde{\Omega}(1/ \sqrt{T})}, addressing an open question posted in \cite{s
yrgkanis2015fast}.
For general m-player games, we show that the swap regret of each player decays a
t rate \tilde{O}(m^{1/2} (n/T)^{3/4}) when they combine optimistic Hedge with th
e classical external-to-internal reduction of Blum and Mansour \cite{blum2007ext
ernal}. The algorithm can also be modified to achieve the same rate against itse
lf and a rate of \tilde{O}(\sqrt{n/T}) against adversaries. Via standard connect
ions, our upper bounds also imply faster convergence to coarse correlated equili
bria in two-player games and to correlated equilibria in multiplayer games.
**********

The Origins and Prevalence of Texture Bias in Convolutional Neural Networks
Katherine Hermann, Ting Chen, Simon Kornblith
Recent work has indicated that, unlike humans, ImageNet-trained CNNs tend to cla
ssify images by texture rather than by shape. How pervasive is this bias, and wh
ere does it come from? We find that, when trained on datasets of images with con
flicting shape and texture, CNNs learn to classify by shape at least as easily a
s by texture. What factors, then, produce the texture bias in CNNs trained on Im
ageNet? Different unsupervised training objectives and different architectures h
ave small but significant and largely independent effects on the level of textur
e bias. However, all objectives and architectures still lead to models that make
 texture-based classification decisions a majority of the time, even if shape in
formation is decodable from their hidden representations. The effect of data aug
mentation is much larger. By taking less aggressive random crops at training tim
e and applying simple, naturalistic augmentation (color distortion, noise, and b
lur), we train models that classify ambiguous images by shape a majority of the
time, and outperform baselines on out-of-distribution test sets. Our results ind
icate that apparent differences in the way humans and ImageNet-trained CNNs proc
ess images may arise not primarily from differences in their internal workings,
but from differences in the data that they see.
**********

Time-Reversal Symmetric ODE Network

In Huh, Eunho Yang, Sung Ju Hwang, Jinwoo Shin
Time-reversal symmetry, which requires that the dynamics of a system should not change with the reversal of time axis, is a fundamental property that frequently holds in classical and quantum mechanics. In this paper, we propose a novel loss function that measures how well our ordinary differential equation (ODE) networks comply with this time-reversal symmetry; it is formally defined by the discrepancy in the time evolutions of ODE networks between forward and backward dynamics. Then, we design a new framework, which we name as Time-Reversal Symmetric ODE Networks (TRS-ODENs), that can learn the dynamics of physical systems more sample-efficiently by learning with the proposed loss function. We evaluate TRS-ODENs on several classical dynamics, and find they can learn the desired time evolution from observed noisy and complex trajectories. We also show that, even for systems that do not possess the full time-reversal symmetry, TRS-ODENs can achieve better predictive performances over baselines.
**********
Provable Overlapping Community Detection in Weighted Graphs
Jimit Majmudar, Stephen Vavasis
Community detection is a widely-studied unsupervised learning problem in which the task is to group similar entities together based on observed pairwise entity interactions. This problem has applications in diverse domains such as social network analysis and computational biology. There is a significant amount of literature studying this problem under the assumption that the communities do not overlap. When the communities are allowed to overlap, often a \textit{pure nodes} assumption is made, i.e. each community has a node that belongs exclusively to that community. This assumption, however, may not always be satisfied in practice. In this paper, we provide a provable method to detect overlapping communities in weighted graphs without explicitly making the pure nodes assumption. Moreover, contrary to most existing algorithms, our approach is based on convex optimization, for which many useful theoretical properties are already known. We demonstrate the success of our algorithm on artificial and real-world datasets.
**********
Fast Unbalanced Optimal Transport on a Tree
Ryoma Sato, Makoto Yamada, Hisashi Kashima
This study examines the time complexities of the unbalanced optimal transport problems from an algorithmic perspective for the first time. We reveal which problems in unbalanced optimal transport can/cannot be solved efficiently. Specifically, we prove that the Kantorovich Rubinstein distance and optimal partial transport in Euclidean metric cannot be computed in strongly subquadratic time under the strong exponential time hypothesis. Then, we propose an algorithm that solves a more general unbalanced optimal transport problem exactly in quasi-linear time on a tree metric. The proposed algorithm processes a tree with one million nodes in less than one second. Our analysis forms a foundation for the theoretical study of unbalanced optimal transport algorithms and opens the door to the applications of unbalanced optimal transport to million-scale datasets.
**********
Acceleration with a Ball Optimization Oracle
Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, Kevin Tian
Consider an oracle which takes a point x and returns the minimizer of a convex function f in an l2 ball of radius r around x. It is straightforward to show that roughly r^{-1}\log(1/epsilon) calls to the oracle suffice to find an \epsilon-approximate minimizer of f in an l2 unit ball. Perhaps surprisingly, this is not optimal: we design an accelerated algorithm which attains an epsilon-approximate minimizer with roughly r^{-2/3} \log(1/epsilon) oracle queries, and give a matching lower bound. Further, we implement ball optimization oracles for functions with a locally stable Hessian using a variant of Newton's method and, in certain cases, stochastic first-order methods. The resulting algorithms apply to a number of problems of practical and theoretical import, improving upon previous results for logistic and
linfinity regression and achieving guarantees comparable to the

state-of-the-art for lp regression.
**********
Avoiding Side Effects By Considering Future Tasks
Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, Shane Legg
Designing reward functions is difficult: the designer has to specify what to do (what it means to complete the task) as well as what not to do (side effects that should be avoided while completing the task). To alleviate the burden on the reward designer, we propose an algorithm to automatically generate an auxiliary reward function that penalizes side effects. This auxiliary objective rewards the ability to complete possible future tasks, which decreases if the agent causes side effects during the current task. The future task reward can also give the agent an incentive to interfere with events in the environment that make future tasks less achievable, such as irreversible actions by other agents. To avoid this interference incentive, we introduce a baseline policy that represents a default course of action (such as doing nothing), and use it to filter out future tasks that are not achievable by default. We formally define interference incentives and show that the future task approach with a baseline policy avoids these incentives in the deterministic case. Using gridworld environments that test for side effects and interference, we show that our method avoids interference and is more effective for avoiding side effects than the common approach of penalizing irreversible actions.
**********
Handling Missing Data with Graph Representation Learning
Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J. Kochenderfer, Jure Leskovec
Machine learning with missing data has been approached in many different ways, including feature imputation where missing feature values are estimated based on observed values and label prediction where downstream labels are learned directly from incomplete data. However, existing imputation models tend to have strong prior assumptions and cannot learn from downstream tasks, while models targeting label predictions often involve heuristics and can encounter scalability issues.
Here we propose GRAPE, a framework for feature imputation as well as label prediction. GRAPE tackles the missing data problem using graph representation, where the observations and features are viewed as two types of nodes in a bipartite graph, and the observed feature values as edges. Under the GRAPE framework, the feature imputation is formulated as an edge-level prediction task and the label prediction as a node-level prediction task. These tasks are then solved with Graph Neural Networks. Experimental results on nine benchmark datasets show that GRAPE yields 20% lower mean absolute error for imputation tasks and 10% lower for label prediction tasks, compared with existing state-of-the-art methods.
**********
Improving Auto-Augment via Augmentation-Wise Weight Sharing
Keyu Tian, Chen Lin, Ming Sun, Luping Zhou, Junjie Yan, Wanli Ouyang
The recent progress on automatically searching augmentation policies has boosted the performance substantially for various tasks.
A key component of automatic augmentation search is the evaluation process for a particular augmentation policy, which is utilized to return reward and usually runs thousands of times.
A plain evaluation process, which includes full model training and validation, would be time-consuming.
To achieve efficiency, many choose to sacrifice evaluation reliability for speed.
In this paper, we dive into the dynamics of augmented training of the model.
This inspires us to design a powerful and efficient proxy task based on the Augmentation-Wise Weight Sharing (AWS) to form a fast yet accurate evaluation process in an elegant way.
Comprehensive analysis verifies the superiority of this approach in terms of effectiveness and efficiency.
The augmentation policies found by our method achieve superior accuracies compared with existing auto-augmentation search methods.

On CIFAR-10, we achieve a top-1 error rate of 1.24%, which is currently the best performing single model without extra training data.
On ImageNet, we get a top-1 error rate of 20.36% for ResNet-50, which leads to 3.34% absolute error rate reduction over the baseline augmentation.
**********
MMA Regularization: Decorrelating Weights of Neural Networks by Maximizing the Minimal Angles
Zhennan Wang, Canqun Xiang, Wenbin Zou, Chen Xu
The strong correlation between neurons or filters can significantly weaken the generalization ability of neural networks. Inspired by the well-known Tammes problem, we propose a novel diversity regularization method to address this issue, which makes the normalized weight vectors of neurons or filters distributed on a hypersphere as uniformly as possible, through maximizing the minimal pairwise angles (MMA). This method can easily exert its effect by plugging the MMA regularization term into the loss function with negligible computational overhead. The MMA regularization is simple, efficient, and effective. Therefore, it can be used as a basic regularization method in neural network training. Extensive experiments demonstrate that MMA regularization is able to enhance the generalization ability of various modern models and achieves considerable performance improvements on CIFAR100 and TinyImageNet datasets. In addition, experiments on face verification show that MMA regularization is also effective for feature learning. Code is available at: https://github.com/wznpub/MMA_Regularization.
**********
HRN: A Holistic Approach to One Class Learning
Wenpeng Hu, Mengyu Wang, Qi Qin, Jinwen Ma, Bing Liu
Existing neural network based one-class learning methods mainly use various forms of auto-encoders or GAN style adversarial training to learn a latent representation of the given one class of data. This paper proposes an entirely different approach based on a novel regularization, called holistic regularization (or H-regularization), which enables the system to consider the data holistically, not to produce a model that biases towards some features.  Combined with a proposed 2-norm instance-level data normalization, we obtain an effective one-class learning method, called HRN. To our knowledge, the proposed regularization and the normalization method have not been reported before. Experimental evaluation using both benchmark image classification and traditional anomaly detection datasets show that HRN markedly outperforms the state-of-the-art existing deep/non-deep learning models.
**********
The Generalized Lasso with Nonlinear Observations and Generative Priors
Zhaoqiang Liu, 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.
**********
Fair regression via plug-in estimator and recalibration with statistical guarantees
Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, Massimiliano Pontil
We study the problem of learning an optimal regression function subject to a fairness constraint. It requires that, conditionally on the sensitive feature, the distribution of the function output remains the same. This constraint naturally extends the notion of demographic parity, often used in classification, to the regression setting. We tackle this problem by leveraging on a proxy-discretized version, for which we derive an explicit expression of the optimal fair predictor. This result naturally suggests a two stage approach, in which we first estimate the (unconstrained) regression function from a set of labeled data and then we recalibrate it with another set of unlabeled data. The recalibration step can be efficiently performed via a smooth optimization. We derive rates of convergence of the proposed estimator to the optimal fair predictor both in terms of the r

isk and fairness constraint. Finally, we present numerical experiments illustrat ing that the proposed method is often superior or competitive with state-of-the- art methods.

**********

Modeling Shared responses in Neuroimaging Studies through MultiView ICA

Hugo Richard, Luigi Gresele, Aapo Hyvarinen, Bertrand Thirion, Alexandre Gramfor t, Pierre Ablin

Group studies involving large cohorts of subjects are important to draw general conclusions about brain functional organization.  However, the aggregation of da ta coming from multiple subjects is challenging, since it requires accounting fo r large variability in anatomy, functional topography and stimulus response acro ss individuals. Data modeling is especially hard for ecologically relevant condi tions such as movie watching, where the experimental setup does not imply well-d efined cognitive operations. We propose a novel MultiView Independent Component Analysis (ICA) model for group studies, where data from each subject are modeled as a linear combination of shared independent sources plus noise. Contrary to m ost group-ICA procedures, the likelihood of the model is available in closed for m. We develop an alternate quasi-Newton method for maximizing the likelihood, wh ich is robust and converges quickly. We demonstrate the usefulness of our approa ch first on fMRI data, where our model demonstrates improved sensitivity in iden tifying common sources among subjects. Moreover, the sources recovered by our mo del exhibit lower between-sessions variability than other methods. On magnetoenc ephalography (MEG) data, our method yields more accurate source localization on phantom data. Applied on 200 subjects from the Cam-CAN dataset, it reveals a cle ar sequence of evoked activity in sensor and source space.

**********

Efficient Planning in Large MDPs with Weak Linear Function Approximation

Roshan Shariff, Csaba Szepesvari

Large-scale Markov decision processes (MDPs) require planning algorithms with runtime independent of the number of states of the MDP. We consider the planning problem in MDPs using linear value function approximation with only weak requirements: low approximation error for the optimal value function, and a smal l set of “core” states whose features span those of other states. In particular, w e make no assumptions about the representability of policies or value functions of non-optimal policies. Our algorithm produces almost-optimal actions for any stat e using a generative oracle (simulator) for the MDP, while its computation time sc ales polynomially with the number of features, core states, and actions and the effec tive horizon.

**********

Efficient Learning of Generative Models via Finite-Difference Score Matching

Tianyu Pang, Kun Xu, Chongxuan LI, Yang Song, Stefano Ermon, Jun Zhu

Several machine learning applications involve the optimization of higher-order d erivatives (e.g., gradients of gradients) during training, which can be expensiv e with respect to memory and computation even with automatic differentiation. As a typical example in generative modeling, score matching~(SM) involves the opti mization of the trace of a Hessian. To improve computing efficiency, we rewrite the SM objective and its variants in terms of directional derivatives, and prese nt a generic strategy to efficiently approximate any-order directional derivativ e with finite difference~(FD). Our approximation only involves function evaluati ons, which can be executed in parallel, and no gradient computations. Thus, it r educes the total computational cost while also improving numerical stability. We provide two instantiations by reformulating variants of SM objectives into the FD forms. Empirically, we demonstrate that our methods produce results comparabl e to the gradient-based counterparts while being much more computationally effic ient.

**********
Semialgebraic Optimization for Lipschitz Constants of ReLU Networks
Tong Chen, Jean B. Lasserre, Victor Magron, Edouard Pauwels
The Lipschitz constant of a network plays an important role in many applications of deep learning, such as robustness certification and Wasserstein Generative Adversarial Network. We introduce a semidefinite programming hierarchy to estimate the global and local Lipschitz constant of a multiple layer deep neural network. The novelty is to combine a polynomial lifting for ReLU functions derivatives with a weak generalization of Putinar's positivity certificate. This idea could also apply to other, nearly sparse, polynomial optimization problems in machine learning. We empirically demonstrate that our method provides a trade-off with respect to state of the art linear programming approach, and in some cases we obtain better bounds in less time.
**********
Linear-Sample Learning of Low-Rank Distributions
Ayush Jain, 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.
**********
Transferable Calibration with Lower Bias and Variance in Domain Adaptation
Ximei Wang, Mingsheng Long, Jianmin Wang, Michael Jordan
Domain Adaptation (DA) enables transferring a learning machine from a labeled source domain to an unlabeled target one. While remarkable advances have been made, most of the existing DA methods focus on improving the target accuracy at inference. How to estimate the predictive uncertainty of DA models is vital for decision-making in safety-critical scenarios but remains the boundary to explore. In this paper, we delve into the open problem of Calibration in DA, which is extremely challenging due to the coexistence of domain shift and the lack of target labels. We first reveal the dilemma that DA models learn higher accuracy at the expense of well-calibrated probabilities. Driven by this finding, we propose Transferable Calibration (TransCal) to achieve more accurate calibration with lower bias and variance in a unified hyperparameter-free optimization framework. As a general post-hoc calibration method, TransCal can be easily applied to recalibrate existing DA methods. Its efficacy has been justified both theoretically and empirically.
**********
Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics
Taiji Suzuki
We introduce a new theoretical framework to analyze deep learning optimization with connection to its generalization error.
Existing frameworks such as mean field theory and neural tangent kernel theory for neural network optimization analysis
typically require taking limit of infinite width of the network to show its global convergence.
This potentially makes it difficult to directly deal with finite width network; especially in the neural tangent kernel regime, we cannot reveal favorable properties of neural networks {\it beyond kernel methods}.
To realize more natural analysis, we consider a completely different approach in which
we formulate the parameter training as a transportation map estimation and show its global convergence via the theory of the {\it infinite dimensional Langevin dynamics}.
This enables us to analyze narrow and wide networks in a unifying manner.
Moreover, we give generalization gap and excess risk bounds for the solution obtained by the dynamics.
The excess risk bound achieves the so-called fast learning rate.
In particular, we show an exponential convergence for a classification problem a

nd a minimax optimal rate for a regression problem.
**********
Online Bayesian Goal Inference for Boundedly Rational Planning Agents
Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, Vikash Mansinghka
People routinely infer the goals of others by observing their actions over time. Remarkably, we can do so even when those actions lead to failure, enabling us to assist others when we detect that they might not achieve their goals. How might we endow machines with similar capabilities? Here we present an architecture capable of inferring an agent's goals online from both optimal and non-optimal sequences of actions. Our architecture models agents as boundedly-rational planners that interleave search with execution by replanning, thereby accounting for sub-optimal behavior. These models are specified as probabilistic programs, allowing us to represent and perform efficient Bayesian inference over an agent's goals and internal planning processes. To perform such inference, we develop Sequential Inverse Plan Search (SIPS), a sequential Monte Carlo algorithm that exploits the online replanning assumption of these models, limiting computation by incrementally extending inferred plans as new actions are observed. We present experiments showing that this modeling and inference architecture outperforms Bayesian inverse reinforcement learning baselines, accurately inferring goals from both optimal and non-optimal trajectories involving failure and back-tracking, while generalizing across domains with compositional structure and sparse rewards.
**********
BayReL: Bayesian Relational Learning for Multi-omics Data Integration
Ehsan Hajiramezanali, Arman Hasanzadeh, Nick Duffield, Krishna Narayanan, Xiaoning Qian
High-throughput molecular profiling technologies have produced high-dimensional multi-omics data, enabling systematic understanding of living systems at the genome scale. Studying molecular interactions across different data types helps reveal signal transduction mechanisms across different classes of molecules. In this paper, we develop a novel Bayesian representation learning method that infers the relational interactions across multi-omics data types. Our method, Bayesian Relational Learning (BayReL) for multi-omics data integration, takes advantage of a priori known relationships among the same class of molecules, modeled as a graph at each corresponding view, to learn view-specific latent variables as well as a multi-partite graph that encodes the interactions across views. Our experiments on several real-world datasets demonstrate enhanced performance of BayReL in inferring meaningful interactions compared to existing baselines.
**********
Weakly Supervised Deep Functional Maps for Shape Matching
Abhishek Sharma, Maks Ovsjanikov
A variety of deep functional maps have been proposed recently, from fully supervised to totally unsupervised, with a range of loss functions as well as different regularization terms. However, it is still not clear what are minimum ingredients of a deep functional map pipeline and whether such ingredients unify or generalize all recent work on deep functional maps. We show empirically the minimum components for obtaining state-of-the-art results with different loss functions, supervised as well as unsupervised. Furthermore, we propose a novel framework designed for both full-to-full as well as partial to full shape matching that achieves state of the art results on  several benchmark datasets outperforming, even the fully supervised methods. Our code is publicly available at \url{https://github.com/Not-IITian/Weakly-supervised-Functional-map}
**********
Domain Adaptation with Conditional Distribution Matching and Generalized Label Shift
Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, Geoffrey J. Gordon
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Rethinking the Value of Labels for Improving Class-Imbalanced Learning
Yuzhe Yang, Zhi Xu
Real-world data often exhibits long-tailed distributions with heavy class imbalance, posing great challenges for deep recognition models. We identify a persisting dilemma on the value of labels in the context of imbalanced learning: on the one hand, supervision from labels typically leads to better results than its unsupervised counterparts; on the other hand, heavily imbalanced data naturally incurs ''label bias'' in the classifier, where the decision boundary can be drastically altered by the majority classes. In this work, we systematically investigate these two facets of labels. We demonstrate, theoretically and empirically, that class-imbalanced learning can significantly benefit in both semi-supervised and self-supervised manners. Specifically, we confirm that (1) positively, imbalanced labels are valuable: given more unlabeled data, the original labels can be leveraged with the extra data to reduce label bias in a semi-supervised manner, which greatly improves the final classifier; (2) negatively however, we argue that imbalanced labels are not useful always: classifiers that are first pre-trained in a self-supervised manner consistently outperform their corresponding baselines. Extensive experiments on large-scale imbalanced datasets verify our theoretically grounded strategies, showing superior performance over previous state-of-the-arts. Our intriguing findings highlight the need to rethink the usage of imbalanced labels in realistic long-tailed tasks. Code is available at https://github.com/YyzHarry/imbalanced-semi-self.
**********
Provably Robust Metric Learning
Lu Wang, Xuanqing Liu, Jinfeng Yi, Yuan Jiang, Cho-Jui Hsieh
Metric learning is an important family of algorithms for classi■cation and similarity search, but the robustness of learned metrics against small adversarial perturbations is less studied. In this paper, we show that existing metric learning algorithms, which focus on boosting the clean accuracy, can result in metrics that are less robust than the Euclidean distance. To overcome this problem, we propose a novel metric learning algorithm to ■nd a Mahalanobis distance that is robust against adversarial perturbations, and the robustness of the resulting model is certi■able. Experimental results show that the proposed metric learning algorithm improves both certi■ed robust errors and empirical robust errors (errors under adversarial attacks). Furthermore, unlike neural network defenses which usually encounter a trade-off between clean and robust errors, our method does not sacri■ce clean errors compared with previous metric learning methods.
**********
Iterative Deep Graph Learning for Graph Neural Networks: Better and Robust Node Embeddings
Yu Chen, Lingfei Wu, Mohammed Zaki
In this paper, we propose an end-to-end graph learning framework, namely \textbf{I}terative \textbf{D}eep \textbf{G}raph \textbf{L}earning (\alg), for jointly and iteratively learning graph structure and graph embedding. The key rationale of \alg is to learn a better graph structure based on better node embeddings, and vice versa (i.e., better node embeddings based on a better graph structure). Our iterative method dynamically stops when the learned graph structure approaches close enough to the graph optimized for the downstream prediction task. In addition, we cast the graph learning problem as a similarity metric learning problem and leverage adaptive graph regularization for controlling the quality of the learned graph. Finally, combining the anchor-based approximation technique, we further propose a scalable version of \alg, namely \salg, which significantly reduces the time and space complexity of \alg without compromising the performance. Our extensive experiments on nine benchmarks show that our proposed \alg models can consistently outperform or match the state-of-the-art baselines. Furthermore, \alg can be more robust to adversarial graphs and cope with both transductive and inductive learning.
**********
COPT: Coordinated Optimal Transport on Graphs
Yihe Dong, Will Sawin

We introduce COPT, a novel distance metric between graphs defined via an optimiz
ation routine, computing a coordinated pair of optimal transport maps simultaneo
usly. This gives an unsupervised way to learn general-purpose graph representati
on, applicable to both graph sketching and graph comparison. COPT involves simul
taneously optimizing dual transport plans, one between the vertices of two graph
s, and another between graph signal probability distributions. We show theoretic
ally that our method preserves important global structural information on graphs
, in particular spectral information, and analyze connections to existing studie
s. Empirically, COPT outperforms state of the art methods in graph classificatio
n on both synthetic and real datasets.
**********

No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classificatio
n Problems
Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, Christopher Ré
In real-world classification tasks, each class often comprises multiple finer-gr
ained "subclasses." As the subclass labels are frequently unavailable, models tr
ained using only the coarser-grained class labels often exhibit highly variable
performance across different subclasses. This phenomenon, known as hidden strati
fication, has important consequences for models deployed in safety-critical appl
ications such as medicine. We propose GEORGE, a method to both measure and mitig
ate hidden stratification even when subclass labels are unknown. We first observ
e that unlabeled subclasses are often separable in the feature space of deep mod
els, and exploit this fact to estimate subclass labels for the training data via
 clustering techniques. We then use these approximate subclass labels as a form
of noisy supervision in a distributionally robust optimization objective. We the
oretically characterize the performance of GEORGE in terms of the worst-case gen
eralization error across any subclass. We empirically validate GEORGE on a mix o
f real-world and benchmark image classification datasets, and show that our appr
oach boosts worst-case subclass accuracy by up to 15 percentage points compared
to standard training techniques, without requiring any information about the sub
classes.
**********

Model Rubik’s Cube: Twisting Resolution, Depth and Width for TinyNets
Kai Han, Yunhe Wang, Qiulin Zhang, Wei Zhang, Chunjing XU, Tong Zhang
To obtain excellent deep neural architectures, a series of techniques are carefu
lly designed in EfficientNets. The giant formula for simultaneously enlarging th
e resolution, depth and width provides us a Rubik’s cube for neural networks. So
 that we can find networks with high efficiency and excellent performance by twi
sting the three dimensions. This paper aims to explore the twisting rules for ob
taining deep neural networks with minimum model sizes and computational costs. D
ifferent from the network enlarging, we observe that resolution and depth are mo
re important than width for tiny networks. Therefore, the original method, \ie t
he compound scaling in EfficientNet is no longer suitable. To this end, we summa
rize a tiny formula for downsizing neural architectures through a series of smal
ler models derived from the EfficientNet-B0 with the FLOPs constraint. Experimen
tal results on the ImageNet benchmark illustrate that our TinyNet performs much
better than the smaller version of EfficientNets using the inversed giant formul
a. For instance, our TinyNet-E achieves a 59.9\% Top-1 accuracy with only 24M FL
OPs, which is about 1.9\% higher than that of the previous best MobileNetV3 with
 similar computational cost. Code will be available at \url{https://github.com/h
uawei-noah/CV-Backbones/tree/master/tinynet}, and \url{https://gitee.com/mindspo
re/mindspore/tree/master/model_zoo/research/cv/tinynet}.
**********

Self-Adaptive Training: beyond Empirical Risk Minimization
Lang Huang, Chao Zhang, Hongyang Zhang
We propose self-adaptive training---a new training algorithm that dynamically ca
librates training process by model predictions without incurring extra computati
onal cost---to improve generalization of deep learning for potentially corrupted
 training data. This problem is important to robustly learning from data that ar
e corrupted by, e.g., random noises and adversarial examples. The standard empir

ical risk minimization (ERM) for such data, however, may easily overfit noises a
nd thus suffers from sub-optimal performance. In this paper, we observe that mod
el predictions can substantially benefit the training process: self-adaptive tra
ining significantly mitigates the overfitting issue and improves generalization
over ERM under both random and adversarial noises. Besides, in sharp contrast to
 the recently-discovered double-descent phenomenon in ERM, self-adaptive trainin
g exhibits a single-descent error-capacity curve, indicating that such a phenome
non might be a result of overfitting of noises. Experiments on the CIFAR and Ima
geNet datasets verify the effectiveness of our approach in two applications: cla
ssification with label noise and selective classification.
**********

Effective Dimension Adaptive Sketching Methods for Faster Regularized Least-Squa
res Optimization
Jonathan Lacotte, Mert Pilanci
We propose a new randomized algorithm for solving L2-regularized least-squares p
roblems based on sketching. We consider two of the most popular random embedding
s, namely, Gaussian embeddings and the Subsampled Randomized Hadamard Transform
(SRHT). While current randomized solvers for least-squares optimization prescrib
e an embedding dimension at least greater than the data dimension, we show that
the embedding dimension can be reduced to the effective dimension of the optimiz
ation problem, and still preserve high-probability convergence guarantees. In th
is regard, we derive sharp matrix deviation inequalities over ellipsoids for bot
h Gaussian and SRHT embeddings. Specifically, we improve on the constant of a cl
assical Gaussian concentration bound whereas, for SRHT embeddings, our deviation
 inequality involves a novel technical approach. Leveraging these bounds, we are
 able to design a practical and adaptive algorithm which does not require to kno
w the effective dimension beforehand. Our method starts with an initial embeddin
g dimension equal to 1 and, over iterations, increases the embedding dimension u
p to the effective one at most. Hence, our algorithm improves the state-of-the-a
rt computational complexity for solving regularized least-squares problems. Furt
her, we show numerically that it outperforms standard iterative solvers such as
the conjugate gradient method and its pre-conditioned version on several standar
d machine learning datasets.
**********

Near-Optimal Comparison Based Clustering
Michaël Perrot, Pascal Esser, Debarghya Ghoshdastidar
The goal of clustering is to group similar objects into meaningful partitions. T
his process is well understood when an explicit similarity measure between the o
bjects is given. However, far less is known when this information is not readily
 available and, instead, one only observes ordinal comparisons such as ``object
i is more similar to j than to k.'' In this paper, we tackle this problem using
a two-step procedure: we estimate a pairwise similarity matrix from the comparis
ons before using a clustering method based on semi-definite programming (SDP). W
e theoretically show that our approach can exactly recover a planted clustering
using a near-optimal number of passive comparisons. We empirically validate our
theoretical findings and demonstrate the good behaviour of our method on real da
ta.
**********

Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Me
asurement
Xin Liu, Josh Fromm, Shwetak Patel, Daniel McDuff
Telehealth and remote health monitoring have become increasingly important durin
g the SARS-CoV-2 pandemic and it is widely expected that this will have a lastin
g impact on healthcare practices. These tools can help reduce the risk of exposi
ng patients and medical staff to infection, make healthcare services more access
ible, and allow providers to see more patients. However, objective measurement o
f vital signs is challenging without direct contact with a patient. We present a
 video-based and on-device optical cardiopulmonary vital sign measurement approa
ch. It leverages a novel multi-task temporal shift convolutional attention netwo
rk (MTTS-CAN) and enables real-time cardiovascular and respiratory measurements

on mobile platforms. We evaluate our system on an Advanced RISC Machine (ARM) CP U and achieve state-of-the-art accuracy while running at over 150 frames per sec ond which enables real-time applications. Systematic experimentation on large be nchmark datasets reveals that our approach leads to substantial (20%-50%) reduct ions in error and generalizes well across datasets.
**********
A new convergent variant of Q-learning with linear function approximation
Diogo Carvalho, Francisco S. Melo, Pedro Santos
In this work, we identify a novel set of conditions that ensure convergence with probability 1 of Q-learning with linear function approximation, by proposing a two time-scale variation thereof. In the faster time scale, the algorithm featur es an update similar to that of DQN, where the impact of bootstrapping is attenu ated by using a Q-value estimate akin to that of the target network in DQN. The slower time-scale, in turn, can be seen as a modified target network update. We establish the convergence of our algorithm, provide an error bound and discuss o ur results in light of existing convergence results on reinforcement learning wi th function approximation. Finally, we illustrate the convergent behavior of our method in domains where standard Q-learning has previously been shown to diverg e.
**********
TaylorGAN: Neighbor-Augmented Policy Update Towards Sample-Efficient Natural Lan guage Generation
Chun-Hsing Lin, Siang-Ruei Wu, Hung-yi Lee, Yun-Nung Chen
Score function-based natural language generation (NLG) approaches such as REINFO RCE, in general, suffer from low sample efficiency and training instability prob lems.
This is mainly due to the non-differentiable nature of the discrete space sampli ng and thus these methods have to treat the discriminator as a black box and ign ore the gradient information.
To improve the sample efficiency and reduce the variance of REINFORCE, we propos e a novel approach, TaylorGAN, which augments the gradient estimation by off-pol icy update and the first-order Taylor expansion.
This approach enables us to train NLG models from scratch with smaller batch siz e --- without maximum likelihood pre-training, and outperforms existing GAN-base d methods on multiple metrics of quality and diversity.
**********
Neural Networks with Small Weights and Depth-Separation Barriers
Gal Vardi, 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-auth ors prior to requesting a name change in the electronic proceedings.
**********
Untangling tradeoffs between recurrence and self-attention in artificial neural networks
Giancarlo Kerg, Bhargav Kanuparthi, Anirudh Goyal ALIAS PARTH GOYAL, Kyle Goyett e, Yoshua Bengio, Guillaume Lajoie
Attention and self-attention mechanisms, are now central to state-of-the-art dee p learning on sequential tasks. However, most recent progress hinges on heuristi c approaches with limited understanding of attention's role in model optimizatio n and computation, and rely on considerable memory and computational resources t hat scale poorly. In this work, we present a formal analysis of how self-attenti on affects gradient propagation in recurrent networks, and prove that it mitigat es the problem of vanishing gradients when trying to capture long-term dependenc ies by establishing concrete bounds for gradient norms. Building on these result s, we propose a relevancy screening mechanism, inspired by the cognitive process of memory consolidation, that allows for a scalable use of sparse self-attentio n with recurrence. While providing guarantees to avoid vanishing gradients, we u se simple numerical experiments to demonstrate the tradeoffs in performance and computational resources by efficiently balancing attention and recurrence. Based

on our results, we propose a concrete direction of research to improve scalability of attentive networks.
**********
Dual-Free Stochastic Decentralized Optimization with Variance Reduction
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-authors prior to requesting a name change in the electronic proceedings.
**********
Online Learning in Contextual Bandits using Gated Linear Networks
Eren Sezener, Marcus Hutter, David Budden, Jianan Wang, Joel Veness
We introduce a new and completely online contextual bandit algorithm called Gated Linear Contextual Bandits (GLCB). This algorithm is based on Gated Linear Networks (GLNs), a recently introduced deep learning architecture with properties well-suited to the online setting. Leveraging data-dependent gating properties of the GLN we are able to estimate prediction uncertainty with effectively zero algorithmic overhead. We empirically evaluate GLCB compared to 9 state-of-the-art algorithms that leverage deep neural networks, on a standard benchmark suite of discrete and continuous contextual bandit problems. GLCB obtains mean first-place despite being the only online method, and we further support these results with a theoretical study of its convergence properties.
**********
Throughput-Optimal Topology Design for Cross-Silo Federated Learning
Othmane MARFOQ, CHUAN XU, Giovanni Neglia, Richard Vidal
Federated learning usually employs a client-server architecture where an orchestrator iteratively aggregates model updates from remote clients and pushes them back a refined model. This approach may be inefficient in cross-silo settings, as close-by data silos with high-speed access links may exchange information faster than with the orchestrator, and the orchestrator may become a communication bottleneck.
In this paper we define the problem of topology design for cross-silo federated learning using the theory of max-plus linear systems to compute the system throughput---number of communication rounds per time unit. We also propose practical algorithms that, under the knowledge of measurable network characteristics, find a topology with the largest throughput or with provable throughput guarantees. In realistic Internet networks with 10~Gbps access links for silos, our algorithms speed up training by a factor 9 and 1.5 in comparison to the master-slave architecture and to state-of-the-art MATCHA, respectively. Speedups are even larger with slower access links.
**********
Quantized Variational Inference
Amir Dib
We present Quantized Variational Inference, a new algorithm for Evidence Lower Bound minimization. We show how Optimal Voronoi Tesselation produces variance free gradients for Evidence Lower Bound (ELBO) optimization at the cost of introducing asymptotically decaying bias. Subsequently, we propose a Richardson extrapolation type method to improve this bound. We show that using the Quantized Variational Inference framework leads to fast convergence for both score function and the reparametrized gradient estimator at a comparable computational cost. Finally, we propose several experiments to assess the performance of our method and its limitations.
**********
Asymptotically Optimal Exact Minibatch Metropolis-Hastings
Ruqi Zhang, A. Feder Cooper, Christopher M. De Sa
Metropolis-Hastings (MH) is a commonly-used MCMC algorithm, but it can be intractable on large datasets due to requiring computations over the whole dataset. In this paper, we study \emph{minibatch MH} methods, which instead use subsamples to enable scaling. We observe that most existing minibatch MH methods are inexact (i.e. they may change the target distribution), and show that this inexactness

can cause arbitrarily large errors in inference. We propose a new exact minibatch MH method, \emph{TunaMH}, which exposes a tunable trade-off between its minibatch size and its theoretically guaranteed convergence rate. We prove a lower bound on the batch size that any minibatch MH method \emph{must} use to retain exactness while guaranteeing fast convergence---the first such bound for minibatch MH---and show TunaMH is asymptotically optimal in terms of the batch size. Empirically, we show TunaMH outperforms other exact minibatch MH methods on robust linear regression, truncated Gaussian mixtures, and logistic regression.
**********
Learning Search Space Partition for Black-box Optimization using Monte Carlo Tree Search
Linnan Wang, Rodrigo Fonseca, Yuandong Tian
High dimensional black-box optimization has broad applications but remains a challenging problem to solve. Given a set of samples xi, yi, building a global model (like Bayesian Optimization (BO)) suffers from the curse of dimensionality in the high-dimensional search space, while a greedy search may lead to sub-optimality. By recursively splitting the search space into regions with high/low function values, recent works like LaNAS shows good performance in Neural Architecture Search (NAS), reducing the sample complexity empirically. In this paper, we coin LA-MCTS that extends LaNAS to other domains. Unlike previous approaches, LA-MCTS learns the partition of the search space using a few samples and their function values in an online fashion. While LaNAS uses linear partition and performs uniform sampling in each region, our LA-MCTS adopts a nonlinear decision boundary and learns a local model to pick good candidates. If the nonlinear partition function and the local model fits well with ground-truth black-box function, then good partitions and candidates can be reached with much fewer samples. LA-MCTS serves as a meta-algorithm by using existing black-box optimizers (e.g., BO, TuRBO as its local models, achieving strong performance in general black-box optimization and reinforcement learning benchmarks, in particular for high-dimensional problems.
**********
Feature Shift Detection: Localizing Which Features Have Shifted via Conditional Distribution Tests
Sean Kulinski, Saurabh Bagchi, David I. Inouye
While previous distribution shift detection approaches can identify if a shift has occurred, these approaches cannot localize which specific features have caused a distribution shift---a critical step in diagnosing or fixing any underlying issue. For example, in military sensor networks, users will want to detect when one or more of the sensors has been compromised, and critically, they will want to know which specific sensors might be compromised. Thus, we first define a formalization of this problem as multiple conditional distribution hypothesis tests and propose both non-parametric and parametric statistical tests. For both efficiency and flexibility, we then propose to use a test statistic based on the density model score function (\ie gradient with respect to the input)---which can easily compute test statistics for all dimensions in a single forward and backward pass. Any density model could be used for computing the necessary statistics including deep density models such as normalizing flows or autoregressive models. We additionally develop methods for identifying when and where a shift occurs in multivariate time-series data and show results for multiple scenarios using realistic attack models on both simulated and real-world data.
**********
Unifying Activation- and Timing-based Learning Rules for Spiking Neural Networks
Jinseok Kim, Kyungsu Kim, Jae-Joon Kim
For the gradient computation across the time domain in Spiking Neural Networks (SNNs) training, two different approaches have been independently studied. The first is to compute the gradients with respect to the change in spike activation (activation-based methods), and the second is to compute the gradients with respect to the change in spike timing (timing-based methods). In this work, we present a comparative study of the two methods and propose a new supervised learning method that combines them. The proposed method utilizes each individual spike mor

e effectively by shifting spike timings as in the timing-based methods as well a s generating and removing spikes as in the activation-based methods. Experimenta l results showed that the proposed method achieves higher performance in terms o f both accuracy and efficiency than the previous approaches.
**********

Space-Time Correspondence as a Contrastive Random Walk
Allan Jabri, Andrew Owens, Alexei Efros
This paper proposes a simple self-supervised approach for learning a representat ion for visual correspondence from raw video. We cast correspondence as predicti on of links in a space-time graph constructed from video. In this graph, the nod es are patches sampled from each frame, and nodes adjacent in time can share a d irected edge.  We learn a representation in which pairwise similarity defines tr ansition probability of a random walk, such that prediction of long-range corres pondence is computed as a walk along the graph. We optimize the representation t o place high probability along paths of similarity. Targets for learning are for med without supervision, by cycle-consistency:  the objective is to maximize the  likelihood of returning to the initial node when walking along a graph construc ted from a palindrome of frames. Thus, a single path-level constraint implicitly  supervises chains of intermediate comparisons.  When used as a similarity metri c without adaptation, the learned representation outperforms the self-supervised  state-of-the-art on label propagation tasks involving objects, semantic parts,  and pose. Moreover, we demonstrate that a technique we call edge dropout, as wel l as self-supervised adaptation at test-time, further improve transfer for objec t-centric correspondence.
**********

The Flajolet-Martin Sketch Itself Preserves Differential Privacy: Private Counti ng with Minimal Space
Adam Smith, Shuang Song, 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-auth ors prior to requesting a name change in the electronic proceedings.
**********

Exponential ergodicity of mirror-Langevin diffusions
Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, Austin S tromme
Motivated by the problem of sampling from ill-conditioned log-concave distributi ons, we give a clean non-asymptotic convergence analysis of mirror-Langevin diff usions as introduced in Zhang et al. (2020). As a special case of this framework , we propose a class of diffusions called Newton-Langevin diffusions and prove t hat they converge to stationarity exponentially fast with a rate which not only  is dimension-free, but also has no dependence on the target distribution. We giv e an application of this result to the problem of sampling from the uniform dist ribution on a convex body using a strategy inspired by interior-point methods. O ur general approach follows the recent trend of linking sampling and optimizatio n and highlights the role of the chi-squared divergence. In particular, it yield s new results on the convergence of the vanilla Langevin diffusion in Wasserstei n distance.
**********

An Efficient Framework for Clustered Federated Learning
Avishek Ghosh, Jichan Chung, Dong Yin, Kannan Ramchandran
We address the problem of Federated Learning (FL) where users are distributed an d partitioned into clusters. This setup captures settings where different groups  of users have their own objectives (learning tasks) but by aggregating their da ta with others in the same cluster (same learning task), they can leverage the s trength in numbers in order to perform more efficient Federated Learning. We pro pose a new framework dubbed the Iterative Federated Clustering Algorithm (IFCA),  which alternately estimates the cluster identities of the users and optimizes m odel parameters for the user clusters via gradient descent. We analyze the conve rgence rate of this algorithm first in a linear model with squared loss and then

for generic strongly convex and smooth loss functions. We show that in both settings, with good initialization, IFCA converges at an exponential rate, and discuss the optimality of the statistical error rate. When the clustering structure is ambiguous, we propose to train the models by combining IFCA with the weight sharing technique in multi-task learning. In the experiments, we show that our algorithm can succeed even if we relax the requirements on initialization with random initialization and multiple restarts. We also present experimental results showing that our algorithm is efficient in non-convex problems such as neural networks. We demonstrate the benefits of IFCA over the baselines on several clustered FL benchmarks.
**********
Autoencoders that don't overfit towards the Identity
Harald Steck
Autoencoders (AE) aim to reproduce the output from the input. They may hence tend to overfit towards learning the identity-function between the input and output, i.e., they may predict each feature in the output from itself in the input. This is not useful, however, when AEs are used for prediction tasks in the presence of noise in the data. It may seem intuitively evident that this kind of overfitting is prevented by training a denoising AE, as the dropped-out features have to be predicted from the other features. In this paper, we consider linear autoencoders, as they facilitate analytic solutions, and first show that denoising / dropout actually prevents the overfitting towards the identity-function only to the degree that it is penalized by the induced L2-norm regularization. In the main theorem of this paper, we show that the emphasized denoising AE is indeed capable of completely eliminating the overfitting towards the identity-function. Our derivations reveal several new insights, including the closed-form solution of the full-rank model, as well as a new (near-)orthogonality constraint in the low-rank model. While this constraint is conceptually very different from the regularizers recently proposed, their resulting effects on the learned embeddings are empirically similar. Our experiments on three well-known data-sets corroborate the various theoretical insights derived in this paper.
**********
Polynomial-Time Computation of Optimal Correlated Equilibria in Two-Player Extensive-Form Games with Public Chance Moves and Beyond
Gabriele Farina, Tuomas Sandholm
Unlike normal-form games, where correlated equilibria have been studied for more than 45 years, extensive-form correlation is still generally not well understood. Part of the reason for this gap is that the sequential nature of extensive-form games allows for a richness of behaviors and incentives that are not possible in normal-form settings. This richness translates to a significantly different complexity landscape surrounding extensive-form correlated equilibria. As of today, it is known that finding an optimal extensive-form correlated equilibrium (EFCE), extensive-form coarse correlated equilibrium (EFCCE), or normal-form coarse correlated equilibrium (NFCCE) in a two-player extensive-form game is computationally tractable when the game does not include chance moves, and intractable when the game involves chance moves. In this paper we significantly refine this complexity threshold by showing that, in two-player games, an optimal correlated equilibrium can be computed in polynomial time, provided that a certain condition is satisfied. We show that the condition holds, for example, when all chance moves are public, that is, both players observe all chance moves. This implies that an optimal EFCE, EFCCE and NFCCE can be computed in polynomial time in the game size in two-player games with public chance moves.
**********
Parameterized Explainer for Graph Neural Network
Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, Xiang Zhang
Despite recent progress in Graph Neural Networks (GNNs), explaining predictions made by GNNs remains a challenging open problem.
The leading method mainly addresses the local explanations (i.e., important subgraph structure and node features) to interpret why a GNN model makes the predict

ion for a single instance, e.g. a node or a graph. As a result, the explanation generated is painstakingly customized for each instance. The unique explanation interpreting each instance
independently is not sufficient to provide a global understanding of the learned GNN model, leading to the lack of generalizability and hindering it from being used in the inductive setting. Besides, as it is designed for explaining a single instance, it is challenging to explain a set of instances naturally (e.g., graphs of a given class).
In this study, we address these key challenges and propose PGExplainer, a parameterized explainer for GNNs. PGExplainer adopts a deep neural network to parameterize the generation process of explanations, which enables PGExplainer a natural approach to multi-instance explanations. Compared to the existing work, PGExplainer has a better generalization power and can be utilized in an inductive setting easily. Experiments on both synthetic and real-life datasets show highly competitive performance with up to 24.7\% relative improvement in AUC on explaining graph classification over the leading baseline.
**********

Recursive Inference for Variational Autoencoders
Minyoung Kim, Vladimir Pavlovic
Inference networks of traditional Variational Autoencoders (VAEs) are typically amortized, resulting in relatively inaccurate posterior approximation compared to instance-wise variational optimization. Recent semi-amortized approaches were proposed to address this drawback; however, their iterative gradient update procedures can be computationally demanding. In this paper, we consider a different approach of building a mixture inference model. We propose a novel recursive mixture estimation algorithm for VAEs that iteratively augments the current mixture with new components so as to maximally reduce the divergence between the variational and the true posteriors. Using the functional gradient approach, we devise an intuitive learning criteria for selecting a new mixture component: the new component has to improve the data likelihood (lower bound) and, at the same time, be as divergent from the current mixture distribution as possible, thus increasing representational diversity. Although there have been similar approaches recently, termed boosted variational inference (BVI), our methods differ from BVI in several aspects, most notably that ours deal with recursive inference in VAEs in the form of amortized inference, while BVI is developed within the standard VI framework, leading to a non-amortized single optimization instance, inappropriate for VAEs. A crucial benefit of our approach is that the inference at test time needs a single feed-forward pass through the mixture inference network, making it significantly faster than the semi-amortized approaches. We show that our approach yields higher test data likelihood than the state-of-the-arts on several benchmark datasets.
**********

Flexible mean field variational inference using mixtures of non-overlapping exponential families
Jeffrey Spence
Sparse models are desirable for many applications across diverse domains as they can perform automatic variable selection, aid interpretability, and provide regularization.  When fitting sparse models in a Bayesian framework, however, analytically obtaining a posterior distribution over the parameters of interest is intractable for all but the simplest cases.  As a result practitioners must rely on either sampling algorithms such as Markov chain Monte Carlo or variational methods to obtain an approximate posterior.  Mean field variational inference is a particularly simple and popular framework that is often amenable to analytically deriving closed-form parameter updates.  When all distributions in the model are members of exponential families and are conditionally conjugate, optimization schemes can often be derived by hand.  Yet, I show that using standard mean field variational inference can fail to produce sensible results for models with sparsity-inducing priors, such as the spike-and-slab.  Fortunately, such pathological behavior can be remedied as I show that mixtures of exponential family distributions with non-overlapping support form an exponential family.  In particular,

any mixture of an exponential family of diffuse distributions and a point mass at zero to model sparsity forms an exponential family. Furthermore, specific choices of these distributions maintain conditional conjugacy. I use two applications to motivate these results: one from statistical genetics that has connections to generalized least squares with a spike-and-slab prior on the regression coefficients; and sparse probabilistic principal component analysis. The theoretical results presented here are broadly applicable beyond these two examples.

**********

HYDRA: Pruning Adversarially Robust Neural Networks

Vikash Sehwag, Shiqi Wang, Prateek Mittal, Suman Jana

In safety-critical but computationally resource-constrained applications, deep learning faces two key challenges: lack of robustness against adversarial attacks and large neural network size (often millions of parameters). While the research community has extensively explored the use of robust training and network pruning \emph{independently} to address one of these challenges, only a few recent works have studied them jointly. However, these works inherit a heuristic pruning strategy that was developed for benign training, which performs poorly when integrated with robust training techniques, including adversarial training and verifiable robust training. To overcome this challenge, we propose to make pruning techniques aware of the robust training objective and let the training objective guide the search for which connections to prune. We realize this insight by formulating the pruning objective as an empirical risk minimization problem which is solved efficiently using SGD. We demonstrate that our approach, titled HYDRA, achieves compressed networks with \textit{state-of-the-art} benign and robust accuracy, \textit{simultaneously}. We demonstrate the success of our approach across CIFAR-10, SVHN, and ImageNet dataset with four robust training techniques: iterative adversarial training, randomized smoothing, MixTrain, and CROWN-IBP. We also demonstrate the existence of highly robust sub-networks within non-robust networks.

**********

NVAE: A Deep Hierarchical Variational Autoencoder

Arash Vahdat, Jan Kautz

Normalizing flows, autoregressive models, variational autoencoders (VAEs), and deep energy-based models are among competing likelihood-based frameworks for deep generative learning. Among them, VAEs have the advantage of fast and tractable sampling and easy-to-access encoding networks. However, they are currently outperformed by other models such as normalizing flows and autoregressive models. While the majority of the research in VAEs is focused on the statistical challenges, we explore the orthogonal direction of carefully designing neural architectures for hierarchical VAEs. We propose Nouveau VAE (NVAE), a deep hierarchical VAE built for image generation using depth-wise separable convolutions and batch normalization. NVAE is equipped with a residual parameterization of Normal distributions and its training is stabilized by spectral regularization. We show that NVAE achieves state-of-the-art results among non-autoregressive likelihood-based models on the MNIST, CIFAR-10, CelebA 64, and CelebA HQ datasets and it provides a strong baseline on FFHQ. For example, on CIFAR-10, NVAE pushes the state-of-the-art from 2.98 to 2.91 bits per dimension, and it produces high-quality images on CelebA HQ. To the best of our knowledge, NVAE is the first successful VAE applied to natural images as large as 256x256 pixels. The source code is publicly available.

**********

Can Temporal-Di■erence and Q-Learning Learn Representation? A Mean-Field Theory

Yufeng Zhang, Qi Cai, Zhuoran Yang, Yongxin Chen, Zhaoran Wang

Temporal-di■erence and Q-learning play a key role in deep reinforcement learning, where they are empowered by expressive nonlinear function approximators such as neural networks. At the core of their empirical successes is the learned feature representation, which embeds rich observations, e.g., images and texts, into the latent space that encodes semantic structures. Meanwhile, the evolution of such a feature representation is crucial to the convergence of temporal-di■erence and Q-learning.

**********
What Do Neural Networks Learn When Trained With Random Labels?
Hartmut Maennel, Ibrahim M. Alabdulmohsin, Ilya O. Tolstikhin, Robert Baldock, O
livier Bousquet, Sylvain Gelly, Daniel Keysers
We study deep neural networks (DNNs) trained on natural image data with entirely
 random labels. Despite its popularity in the literature, where it is often used
 to study memorization, generalization, and other phenomena, little is known abo
ut what DNNs learn in this setting. In this paper, we show analytically for conv
olutional and fully connected networks that an alignment between the principal c
omponents of network parameters and data takes place when training with random l
abels. We study this alignment effect by investigating neural networks pre-train
ed on randomly labelled image data  and subsequently fine-tuned on disjoint data
sets with random or real labels. We show how this alignment produces a positive
transfer: networks pre-trained with random labels train faster downstream compar
ed to training from scratch even after accounting for simple effects, such as we
ight scaling. We analyze how competing effects, such as specialization at later
layers, may hide the positive transfer. These effects are studied in several net
work architectures, including VGG16 and ResNet18, on CIFAR10 and ImageNet.
**********
Counterfactual Prediction for Bundle Treatment
Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, Yue He
Estimating counterfactual outcome of different treatments from observational dat
a is an important problem to assist decision making in a variety of fields.
Among the various forms of treatment specification, bundle treatment has been wi
dely adopted in many scenarios, such as recommendation systems and online market
ing.
The bundle treatment usually can be abstracted as a high dimensional binary vect
or, which makes it more challenging for researchers to remove the confounding bi
as in observational data.
In this work, we assume the existence of low dimensional latent structure underl
ying bundle treatment.
Via the learned latent representations of treatments, we propose a novel variati
onal sample re-weighting (VSR) method to eliminate confounding bias by decorrela
ting the treatments and confounders.
Finally, we conduct extensive experiments to demonstrate that the predictive mod
el trained on this re-weighted dataset can achieve more accurate counterfactual
outcome prediction.
**********
Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs
Hongyu Ren, Jure Leskovec
Requests for name changes in the electronic proceedings will be accepted with 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 Disentangled Representations and Group Structure of Dynamical Environme
nts
Robin Quessard, Thomas Barrett, William Clements
Learning disentangled representations is a key step towards effectively discover
ing and modelling the underlying structure of environments. In the natural scien
ces, physics has found great success by describing the universe in terms of symm
etry preserving transformations. Inspired by this formalism, we propose a framew
ork, built upon the theory of group representation, for learning representations
 of a dynamical environment structured around the transformations that generate
its evolution. Experimentally, we learn the structure of explicitly symmetric en
vironments without supervision from observational data generated by sequential i
nteractions. We further introduce an intuitive disentanglement regularisation to
 ensure the interpretability of the learnt representations. We show that our met
hod enables accurate long-horizon predictions, and demonstrate a correlation bet
ween the quality of predictions and disentanglement in the latent space.

**********
Learning Linear Programs from Optimal Decisions
Yingcong Tan, Daria Terekhov, Andrew Delong
We propose a flexible gradient-based framework for learning linear programs from optimal decisions. Linear programs are often specified by hand, using prior knowledge of relevant costs and constraints. In some applications, linear programs must instead be learned from observations of optimal decisions. Learning from optimal decisions is a particularly challenging bilevel problem, and much of the related inverse optimization literature is dedicated to special cases. We tackle the general problem, learning all parameters jointly while allowing flexible parameterizations of costs, constraints, and loss functions. We also address challenges specific to learning linear programs, such as empty feasible regions and non-unique optimal decisions. Experiments show that our method successfully learns synthetic linear programs and minimum-cost multi-commodity flow instances for which previous methods are not directly applicable. We also provide a fast batch-mode PyTorch implementation of the homogeneous interior point algorithm, which supports gradients by implicit differentiation or backpropagation.
**********
Wisdom of the Ensemble: Improving Consistency of Deep Learning Models
Lijing Wang, Dipanjan Ghosh, Maria Gonzalez Diaz, Ahmed Farahat, Mahbubul Alam, Chetan Gupta, Jiangzhuo Chen, Madhav Marathe
Deep learning classifiers are assisting humans in making decisions and hence the user's trust in these models is of paramount importance. Trust is often a function of constant behavior. From an AI model perspective it means given the same input the user would expect the same output, especially for correct outputs, or in other words consistently correct outputs. This paper studies a model behavior in the context of periodic retraining of deployed models where the outputs from successive generations of the models might not agree on the correct labels assigned to the same input. We formally define consistency and correct-consistency of a learning model. We prove that consistency and correct-consistency of an ensemble learner is not less than the average consistency and correct-consistency of individual learners and correct-consistency can be improved with a probability by combining learners with accuracy not less than the average accuracy of ensemble component learners. To validate the theory using three datasets and two state-of-the-art deep learning classifiers we also propose an efficient dynamic snapshot ensemble method and demonstrate its value.
Code for our algorithm is available at https://github.com/christa60/dynens.
**********
Universal Function Approximation on Graphs
Rickard Brüel Gabrielsson
In this work we produce a framework for constructing universal function approximators on graph isomorphism classes. We prove how this framework comes with a collection of theoretically desirable properties and enables novel analysis. We show how this allows us to achieve state-of-the-art performance on four different well-known datasets in graph classification and separate classes of graphs that other graph-learning methods cannot. Our approach is inspired by persistent homology, dependency parsing for NLP, and multivalued functions. The complexity of the underlying algorithm is O(#edges x #nodes) and code is publicly available (https://github.com/bruel-gabrielsson/universal-function-approximation-on-graphs).
**********
Accelerating Reinforcement Learning through GPU Atari Emulation
Steven Dalton, iuri frosio
We introduce CuLE (CUDA Learning Environment), a CUDA port of the Atari Learning Environment (ALE) which is used for the development of deep reinforcement algorithms. CuLE overcomes many limitations of existing CPU-based emulators and scales naturally to multiple GPUs. It leverages GPU parallelization to run thousands of games simultaneously and it renders frames directly on the GPU, to avoid the bottleneck arising from the limited CPU-GPU communication bandwidth. CuLE generates up to 155M frames per hour on a single GPU, a finding previously achieved only through a cluster of CPUs. Beyond highlighting the differences between CPU an

d GPU emulators in the context of reinforcement learning, we show how to leverag e the high throughput of CuLE by effective batching of the training data, and sh ow accelerated convergence for A2C+V-trace. CuLE is available at https://github. com/NVlabs/cule.
**********

EvolveGraph: Multi-Agent Trajectory Prediction with Dynamic Relational Reasoning
Jiachen Li, Fan Yang, Masayoshi Tomizuka, Chiho Choi
Multi-agent interacting systems are prevalent in the world, from purely physical systems to complicated social dynamic systems. In many applications, effective understanding of the situation and accurate trajectory prediction of interactive agents play a significant role in downstream tasks, such as decision making and planning. In this paper, we propose a generic trajectory forecasting framework (named EvolveGraph) with explicit relational structure recognition and predictio n via latent interaction graphs among multiple heterogeneous, interactive agents . Considering the uncertainty of future behaviors, the model is designed to prov ide multi-modal prediction hypotheses. Since the underlying interactions may evo lve even with abrupt changes, and different modalities of evolution may lead to different outcomes, we address the necessity of dynamic relational reasoning and adaptively evolving the interaction graphs. We also introduce a double-stage tr aining pipeline which not only improves training efficiency and accelerates conv ergence, but also enhances model performance. The proposed framework is evaluate d on both synthetic physics simulations and multiple real-world benchmark datase ts in various areas. The experimental results illustrate that our approach achie ves state-of-the-art performance in terms of prediction accuracy.
**********

Comparator-Adaptive Convex Bandits
Dirk van der Hoeven, Ashok Cutkosky, Haipeng Luo
We study bandit convex optimization methods that adapt to the norm of the compar ator, a topic that has only been studied before for its full-information counter part. Specifically, we develop convex bandit algorithms with regret bounds that are small whenever the norm of the comparator is small.
We first use techniques from the full-information setting to develop comparator- adaptive algorithms for linear bandits. Then, we extend the ideas to convex band its with Lipschitz or smooth loss functions, using a new single-point gradient e stimator and carefully designed surrogate losses.
**********

Model-based Reinforcement Learning for Semi-Markov Decision Processes with Neura l ODEs
Jianzhun Du, Joseph Futoma, Finale Doshi-Velez
We present two elegant solutions for modeling continuous-time dynamics, in a nov el model-based reinforcement learning (RL) framework for semi-Markov decision pr ocesses (SMDPs), using neural ordinary differential equations (ODEs). Our models accurately characterize continuous-time dynamics and enable us to develop high- performing policies using a small amount of data. We also develop a model-based approach for optimizing time schedules to reduce interaction rates with the envi ronment while maintaining the near-optimal performance, which is not possible fo r model-free methods. We experimentally demonstrate the efficacy of our methods across various continuous-time domains.
**********

The Adaptive Complexity of Maximizing a Gross Substitutes Valuation
Ron Kupfer, Sharon Qian, Eric Balkanski, Yaron Singer
In this paper, we study the adaptive complexity of maximizing a monotone gross s ubstitutes function under a cardinality constraint. Our main result is an algori thm that achieves a 1-epsilon approximation in O(log n) adaptive rounds for any constant epsilon > 0, which is an exponential speedup in parallel running time c ompared to previously studied algorithms for gross substitutes functions. We sho w that the algorithmic results are tight in the sense that there is no algorithm that obtains a constant factor approximation in o(log n) rounds. Both the upper and lower bounds are under the assumption that queries are only on feasible set s (i.e., of size at most k). We also show that under a stronger model, where non

-feasible queries are allowed, there is no non-adaptive algorithm that obtains a n approximation better than 1/2 + epsilon. Both lower bounds extend to the class of OXS functions. Additionally, we conduct experiments on synthetic and real da ta sets to demonstrate the near-optimal performance and efficiency of the algori thm in practice.
**********

A Robust Functional EM Algorithm for Incomplete Panel Count Data
Alexander Moreno, Zhenke Wu, Jamie Roslyn Yap, Cho Lam, David Wetter, Inbal Nahu m-Shani, Walter Dempsey, James M. Rehg
Panel count data describes aggregated counts of recurrent events observed at dis crete time points. To understand dynamics of health behaviors and predict futur e negative events, the field of quantitative behavioral research has evolved to increasingly rely upon panel count data collected via multiple self reports, for example, about frequencies of smoking using in-the-moment surveys on mobile dev ices. However, missing reports are common and present a major barrier to downstr eam statistical learning. As a first step, under a missing completely at random assumption (MCAR), we propose a simple yet widely applicable functional EM algor ithm to estimate the counting process mean function, which is of central interes t to behavioral scientists. The proposed approach wraps several popular panel co unt inference methods, seamlessly deals with incomplete counts and is robust to misspecification of the Poisson process assumption. Theoretical analysis of the proposed algorithm provides finite-sample guarantees by extending parametric E M theory to the general non-parametric setting. We illustrate the utility of the proposed algorithm through numerical experiments and an analysis of smoking ces sation data. We also discuss useful extensions to address deviations from the MC AR assumption and covariate effects.
**********

Graph Stochastic Neural Networks for Semi-supervised Learning
Haibo Wang, Chuan Zhou, Xin Chen, Jia Wu, Shirui Pan, Jilong Wang
Graph Neural Networks (GNNs) have achieved remarkable performance in the task of the semi-supervised node classification. However, most existing models learn a deterministic classification function, which lack sufficient flexibility to expl ore better choices in the presence of kinds of imperfect observed data such as the scarce labeled nodes and noisy graph structure. To improve the rigidness an d inflexibility of deterministic classification functions, this paper proposes a novel framework named Graph Stochastic Neural Networks (GSNN), which aims to mo del the uncertainty of the classification function by simultaneously learning a family of functions, i.e., a stochastic function. Specifically, we introduce a learnable graph neural network coupled with a high-dimensional latent variable to model the distribution of the classification function, and further adopt the amortised variational inference to approximate the intractable joint posterior f or missing labels and the latent variable. By maximizing the lower-bound of the likelihood for observed node labels, the instantiated models can be trained in a n end-to-end manner effectively. Extensive experiments on three real-world datas ets show that GSNN achieves substantial performance gain in different scenarios compared with stat-of-the-art baselines.
**********

Compositional Zero-Shot Learning via Fine-Grained Dense Feature Composition
Dat Huynh, Ehsan Elhamifar
We develop a novel generative model for zero-shot learning to recognize fine-gra ined unseen classes without training samples. Our observation is that generating holistic features of unseen classes fails to capture every attribute needed to distinguish small differences among classes. We propose a feature composition fr amework that learns to extract attribute-based features from training samples an d combines them to construct fine-grained features for unseen classes. Feature c omposition allows us to not only selectively compose features of unseen classes from only relevant training samples, but also obtain diversity among composed fe atures via changing samples used for composition. In addition, instead of buildi ng a global feature of an unseen class, we use all attribute-based features to f orm a dense representation consisting of fine-grained attribute details. To reco

gnize unseen classes, we propose a novel training scheme that uses a discriminat ive model to construct features that are subsequently used to train itself. Ther efore, we directly train the discriminative model on composed features without l earning separate generative models. We conduct experiments on four popular datas ets of DeepFashion, AWA2, CUB, and SUN, showing that our method significantly im proves the state of the art.
**********

A Benchmark for Systematic Generalization in Grounded Language Understanding
Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, Brenden M. Lake
Humans easily interpret expressions that describe unfamiliar situations composed from familiar parts ("greet the pink brontosaurus by the ferris wheel"). Modern neural networks, by contrast, struggle to interpret novel compositions. In this paper, we introduce a new benchmark, gSCAN, for evaluating compositional genera lization in situated language understanding. Going beyond a related benchmark th at focused on syntactic aspects of generalization, gSCAN defines a language grou nded in the states of a grid world, facilitating novel evaluations of acquiring linguistically motivated rules. For example, agents must understand how adjectiv es such as 'small' are interpreted relative to the current world state or how ad verbs such as 'cautiously' combine with new verbs. We test a strong multi-modal baseline model and a state-of-the-art compositional method finding that, in most cases, they fail dramatically when generalization requires systematic compositi onal rules.
**********

Weston-Watkins Hinge Loss and Ordered Partitions
Yutong Wang, Clayton Scott
Multiclass extensions of the support vector machine (SVM) have been formulated i n a variety of ways. A recent empirical comparison of nine such formulations [Do ■an et al. 2016] recommends the variant proposed by Weston and Watkins (WW), des pite the fact that the WW-hinge loss is not calibrated with respect to the 0-1 l oss.
In this work we introduce a novel discrete loss function for multiclass classifi cation, the ordered partition loss, and prove that the WW-hinge loss is calibrat ed with respect to this loss. We also argue that the ordered partition loss is m inimally emblematic among discrete losses satisfying this property. Finally, we apply our theory to justify the empirical observation made by Do■an et al that t he WW-SVM can work well even under massive label noise, a challenging setting fo r multiclass SVMs.
**********

Reinforcement Learning with Augmented Data
Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, Aravind Srini vas
Learning from visual observations is a fundamental yet challenging problem in Re inforcement Learning (RL). Although algorithmic advances combined with convoluti onal neural networks have proved to be a recipe for success, current methods are still lacking on two fronts: (a) data-efficiency of learning and (b) generaliza tion to new environments. To this end, we present Reinforcement Learning with Au gmented Data (RAD), a simple plug-and-play module that can enhance most RL algor ithms. We perform the first extensive study of general data augmentations for RL on both pixel-based and state-based inputs, and introduce two new data augmenta tions - random translate and random amplitude scale. We show that augmentations such as random translate, crop, color jitter, patch cutout, random convolutions, and amplitude scale can enable simple RL algorithms to outperform complex state -of-the-art methods across common benchmarks. RAD sets a new state-of-the-art in terms of data-efficiency and final performance on the DeepMind Control Suite be nchmark for pixel-based control as well as OpenAI Gym benchmark for state-based control. We further demonstrate that RAD significantly improves test-time genera lization over existing methods on several OpenAI ProcGen benchmarks.
**********

Towards Minimax Optimal Reinforcement Learning in Factored Markov Decision Proce sses

Yi Tian, Jian Qian, Suvrit Sra
We study minimax optimal reinforcement learning in episodic factored Markov decision processes (FMDPs), which are MDPs with conditionally independent transition components. Assuming the factorization is known, we propose two model-based algorithms. The first one achieves minimax optimal regret guarantees for a rich class of factored structures, while the second one enjoys better computational complexity with a slightly worse regret. A key new ingredient of our algorithms is the design of a bonus term to guide exploration. We complement our algorithms by presenting several structure dependent lower bounds on regret for FMDPs that reveal the difficulty hiding in the intricacy of the structures.
**********

Graduated Assignment for Joint Multi-Graph Matching and Clustering with Application to Unsupervised Graph Matching Network Learning
Runzhong Wang, Junchi Yan, Xiaokang Yang
This paper considers the setting of jointly matching and clustering multiple graphs belonging to different groups, which naturally rises in many realistic problems. Both graph matching and clustering are challenging (NP-hard) and a joint solution is appealing due to the natural connection of the two tasks. In this paper, we resort to a graduated assignment procedure for soft matching and clustering over iterations, whereby the two-way constraint and clustering confidence are modulated by two separate annealing parameters, respectively. Our technique can be further utilized for end-to-end learning whose loss refers to the cross-entropy between two lines of matching pipelines, as such the keypoint feature extraction CNNs can be learned without ground-truth supervision. Experimental results on real-world benchmarks show our method outperforms learning-free algorithms and performs comparatively against two-graph based supervised graph matching approaches.
**********

Estimating Training Data Influence by Tracing Gradient Descent
Garima Pruthi, Frederick Liu, Satyen Kale, Mukund Sundararajan
We introduce a method called TracIn that computes the influence of a training example on a prediction made by the model. The idea is to trace how the loss on the test point changes during the training process whenever the training example of
interest was utilized. We provide a scalable implementation of TracIn via: (a) a first-order gradient approximation to the exact computation, (b) saved checkpoints
of standard training procedures, and (c) cherry-picking layers of a deep neural network. In contrast with previously proposed methods, TracIn is simple to implement; all it needs is the ability to work with gradients, checkpoints, and loss
functions. The method is general. It applies to any machine learning model trained
using stochastic gradient descent or a variant of it, agnostic of architecture, domain
and task. We expect the method to be widely useful within processes that study and improve training data.
**********

Joint Policy Search for Multi-agent Collaboration with Imperfect Information
Yuandong Tian, Qucheng Gong, Yu Jiang
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  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 Bandits with Corruptions: Regret Lower Bound and No-regret Algorithm
lin yang, Mohammad Hajiesmaili, Mohammad Sadegh Talebi, John C. S. Lui, Wing Shing Wong
This paper studies adversarial bandits with corruptions. In the basic adversarial bandit setting, the reward of arms is predetermined by an adversary who is obl

ivious to the learner’s policy. In this paper, we consider an extended setting i n which an attacker sits in-between the environment and the learner, and is endo wed with a limited budget to corrupt the reward of the selected arm. We have two main results. First, we derive a lower bound on the regret of any bandit algori thm that is aware of the budget of the attacker. Also, for budget-agnostic algor ithms, we characterize an impossibility result demonstrating that even when the attacker has a sublinear budget, i.e., a budget growing sublinearly with time ho rizon T, they fail to achieve a sublinear regret.
Second, we propose ExpRb, a bandit algorithm that incorporates a biased estimato r and a robustness parameter to deal with corruption. We characterize the regret of ExpRb as a function of the corruption budget and show that for the case of a known corruption budget, the regret of ExpRb is tight.
**********

Beta R-CNN: Looking into Pedestrian Detection from Another Perspective
Zixuan Xu, Banghuai Li, Ye Yuan, Anhong Dang
Recently significant progress has been made in pedestrian detection, but it rema ins challenging to achieve high performance in occluded and crowded scenes. It c ould be mostly attributed to the widely used representation of pedestrians, i.e. , 2Daxis-aligned bounding box, which just describes the approximate location and size of the object. Bounding box models the object as a uniform distribution wi thin the boundary, making pedestrians indistinguishable in occluded and crowded scenes due to much noise. To eliminate the problem, we propose a novel represent ation based on 2D beta distribution, named Beta Representation. It pictures a pe destrianby explicitly constructing the relationship between full-body and visibl e boxes, and emphasizes the center of visual mass by assigning different probabi lity valuesto pixels. As a result, Beta Representation is much better for distinguishing highly-overlapped instances in crowded scenes with a new NM S strategy named BetaNMS. What’s more, to fully exploit Beta Representation, a n ovel pipeline Beta R-CNN equipped with BetaHead and BetaMask is proposed, leadin g to high detection performance in occluded and crowded scenes.
**********

Batch Normalization Biases Residual Blocks Towards the Identity Function in Deep Networks
Soham De, Sam Smith
Batch normalization dramatically increases the largest trainable depth of residu al networks, and this benefit has been crucial to the empirical success of deep residual networks on a wide range of benchmarks. We show that this key benefit a rises because, at initialization, batch normalization downscales the residual br anch relative to the skip connection, by a normalizing factor on the order of th e square root of the network depth. This ensures that, early in training, the fu nction computed by normalized residual blocks in deep networks is close to the i dentity function (on average). We use this insight to develop a simple initializ ation scheme that can train deep residual networks without normalization. We als o provide a detailed empirical study of residual networks, which clarifies that, although batch normalized networks can be trained with larger learning rates, t his effect is only beneficial in specific compute regimes, and has minimal benef its when the batch size is small.
**********

Learning Retrospective Knowledge with Reverse Reinforcement Learning
Shangtong Zhang, Vivek Veeriah, Shimon Whiteson
We present a Reverse Reinforcement Learning (Reverse RL) approach for representi ng retrospective knowledge. General Value Functions (GVFs) have enjoyed great su ccess in representing predictive knowledge, i.e., answering questions about poss ible future outcomes such as “how much fuel will be consumed in expectation if w e drive from A to B?”. GVFs, however, cannot answer questions like “how much fue l do we expect a car to have given it is at B at time t?”. To answer this questi on, we need to know when that car had a full tank and how that car came to B. Si nce such questions emphasize the influence of possible past events on the presen t, we refer to their answers as retrospective knowledge. In this paper, we show how to represent retrospective knowledge with Reverse GVFs, which are trained vi

a Reverse RL. We demonstrate empirically the utility of Reverse GVFs in both rep
resentation learning and anomaly detection.
**********

Dialog without Dialog Data: Learning Visual Dialog Agents from VQA Data
Michael Cogswell, Jiasen Lu, Rishabh Jain, Stefan Lee, Devi Parikh, Dhruv Batra
Can we develop visually grounded dialog agents that can efficiently adapt to new
 tasks without forgetting how to talk to people? Such agents could leverage a la
rger variety of existing data to generalize to a new task, minimizing expensive
data collection and annotation. In this work, we study a setting we call
"Dialog without Dialog", which requires agents to develop visually grounded dial
og models that can adapt to new tasks without language level supervision.
By factorizing intention and language, our model minimizes linguistic drift afte
r fine-tuning for new tasks. We present qualitative results, automated metrics,
and human studies that all show our model can adapt to new tasks and maintain la
nguage quality. Baselines either fail to perform well at new tasks or experience
 language drift, becoming unintelligible to humans. Code has been made available
 at: https://github.com/mcogswell/dialogwithoutdialog.
**********

GCOMB: Learning Budget-constrained Combinatorial Algorithms over Billion-sized G
raphs
Sahil Manchanda, AKASH MITTAL, Anuj Dhawan, Sourav Medya, Sayan Ranu, Ambuj Sing
h
There has been an increased interest in discovering heuristics for combinatorial
 problems on graphs through machine learning. While existing techniques have pri
marily focused on obtaining high-quality solutions, scalability to billion-sized
 graphs has not been adequately addressed. In addition, the impact of a budget-c
onstraint, which is necessary for many practical scenarios, remains to be studie
d. In this paper, we propose a framework called GCOMB to bridge these gaps. GCOM
B trains a Graph Convolutional Network (GCN) using a novel probabilistic greedy
mechanism to predict the quality of a node. To further facilitate the combinator
ial nature of the problem, GCOMB utilizes a Q-learning framework, which is made
efficient through importance sampling. We perform extensive experiments on real
graphs to benchmark the efficiency and efficacy of GCOMB. Our results establish
that GCOMB is 100 times faster and marginally better in quality than state-of-th
e-art algorithms for learning combinatorial algorithms. Additionally, a case-stu
dy on the practical combinatorial problem of Influence Maximization (IM) shows G
COMB is 150 times faster than the specialized IM algorithm IMM with similar qual
ity.
**********

A General Large Neighborhood Search Framework for Solving Integer Linear Program
s
Jialin Song, ravi lanka, Yisong Yue, Bistra Dilkina
This paper studies how to design abstractions of large-scale combinatorial optim
ization problems that can leverage existing state-of-the-art solvers in general-
purpose ways, and that are amenable to data-driven design.  The goal is to arriv
e at new approaches that can reliably outperform existing solvers in wall-clock
time.  We focus on solving integer programs and ground our approach in the large
 neighborhood search (LNS) paradigm, which iteratively chooses a subset of varia
bles to optimize while leaving the remainder fixed.  The appeal of LNS is that i
t can easily use any existing solver as a subroutine, and thus can inherit the b
enefits of carefully engineered heuristic approaches and their software implemen
tations.  We also show that one can learn a good neighborhood selector from trai
ning data.  Through an extensive empirical validation, we demonstrate that our L
NS framework can significantly outperform, in wall-clock time, compared to state
-of-the-art commercial solvers such as Gurobi.
**********

A Theoretical Framework for Target Propagation
Alexander Meulemans, Francesco Carzaniga, Johan Suykens, João Sacramento, Benjam
in F. Grewe
The success of deep learning, a brain-inspired form of AI, has sparked interest

in understanding how the brain could similarly learn across multiple layers of n
eurons. However, the majority of biologically-plausible learning algorithms have
 not yet reached the performance of backpropagation (BP), nor are they built on
strong theoretical foundations. Here, we analyze target propagation (TP), a popu
lar but not yet fully understood alternative to BP, from the standpoint of mathe
matical optimization. Our theory shows that TP is closely related to Gauss-Newto
n optimization and thus substantially differs from BP. Furthermore, our analysis
 reveals a fundamental limitation of difference target propagation (DTP), a well
-known variant of TP, in the realistic scenario of non-invertible neural network
s. We provide a first solution to this problem through a novel reconstruction lo
ss that improves feedback weight training, while simultaneously introducing arch
itectural flexibility by allowing for direct feedback connections from the outpu
t to each hidden layer. Our theory is corroborated by experimental results that
show significant improvements in performance and in the alignment of forward wei
ght updates with loss gradients, compared to DTP.
**********

OrganITE: Optimal transplant donor organ offering using an individual treatment
effect

Jeroen Berrevoets, James Jordon, Ioana Bica, alexander gimson, Mihaela van der S
chaar

Transplant-organs are a scarce medical resource. The uniqueness of each organ an
d the patients' heterogeneous responses to the organs present a unique and chall
enging machine learning problem. In this problem there are two key challenges: (
i) assigning each organ "optimally" to a patient in the queue; (ii) accurately e
stimating the potential outcomes associated with each patient and each possible
organ. In this paper, we introduce OrganITE, an organ-to-patient assignment meth
odology that assigns organs based not only on its own estimates of the potential
 outcomes but also on organ scarcity. By modelling and accounting for organ scar
city we significantly increase total life years across the population, compared
to the existing greedy approaches that simply optimise life years for the curren
t organ available. Moreover, we propose an individualised treatment effect model
 capable of addressing the high dimensionality of the organ space. We test our m
ethod on real and simulated data, resulting in as much as an additional year of
life expectancy as compared to existing organ-to-patient policies.
**********

The Complete Lasso Tradeoff Diagram

Hua Wang, Yachong Yang, Zhiqi Bu, Weijie Su

A fundamental problem in high-dimensional regression is to understand the tradeo
ff between type I and type II errors or, equivalently, false discovery rate (FDR
) and power in variable selection. To address this important problem, we offer t
he first complete diagram that distinguishes all pairs of FDR and power that can
 be asymptotically realized by the Lasso from the remaining pairs, in a regime o
f linear sparsity under random designs. The tradeoff between the FDR and power c
haracterized by our diagram holds no matter how strong the signals are. In parti
cular, our results complete the earlier Lasso tradeoff diagram in previous liter
ature by recognizing two simple constraints on the pairs of FDR and power. The i
mprovement is more substantial when the regression problem is above the Donoho-T
anner phase transition. Finally, we present extensive simulation studies to conf
irm the sharpness of the complete Lasso tradeoff diagram.
**********

On the universality of deep learning

Emmanuel Abbe, Colin Sandon

This paper shows that deep learning, i.e., neural networks trained by SGD, can l
earn in polytime any function class that can be learned in polytime by some algo
rithm, including parities. This universal result is further shown to be robust,
i.e., it holds under possibly poly-noise on the gradients, which gives a separat
ion between deep learning and statistical query algorithms, as the latter are no
t comparably universal due to cases like parities. This also shows that SGD-base
d deep learning does not suffer from the limitations of the perceptron discussed
 by Minsky-Papert '69. The paper further complement this result with a lower-bou

nd on the generalization error of descent algorithms, which implies in particula
r that the robust universality breaks down if the gradients are averaged over la
rge enough batches of samples as in full-GD, rather than fewer samples as in SGD
.
**********

Regression with reject option and application to kNN
Ahmed Zaoui, Christophe Denis, Mohamed Hebiri
We investigate the problem of regression where one is allowed to abstain from pr
edicting. We refer to this framework as regression with reject option as an exte
nsion of classification with reject option. In this context, we focus on the cas
e where the rejection rate is fixed and derive the optimal rule which relies on
thresholding the conditional variance function. We provide a semi-supervised est
imation procedure of the optimal rule involving two datasets: a first labeled da
taset is used to estimate both regression function and conditional variance func
tion while a second unlabeled dataset is exploited to calibrate the desired reje
ction rate. The resulting predictor with reject option is shown to be almost as
good as the optimal predictor with reject option both in terms of risk and rejec
tion rate. We additionally apply our methodology with kNN algorithm and establis
h rates of convergence for the resulting kNN predictor under mild conditions. Fi
nally, a numerical study is performed to illustrate the benefit of using the pro
posed procedure.
**********

The Primal-Dual method for Learning Augmented Algorithms
Etienne Bamas, Andreas Maggiori, Ola Svensson
The extension of classical online algorithms when provided with predictions is
a new and active research area. In this paper, we extend the primal-dual method
for online algorithms in order to incorporate predictions that advise the online
algorithm about the next action to take. We use this framework to obtain novel
algorithms for a variety of online covering problems. We compare our algorithms
to the cost of the true and predicted offline optimal solutions and show that th
ese
algorithms outperform any online algorithm when the prediction is accurate while
maintaining good guarantees when the prediction is misleading.
**********

FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs
Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, Wen Sun
In order to deal with the curse of dimensionality in reinforcement learning (RL)
, it is common practice to make parametric assumptions where values or policies
are functions of some low dimensional feature space. This work focuses on the re
presentation learning question: how can we learn such features? Under the assump
tion that the underlying (unknown) dynamics correspond to a low rank transition
matrix, we show how the representation learning question is related to a particu
lar non-linear matrix decomposition problem. Structurally, we make precise conne
ctions between these low rank MDPs and latent variable models, showing how they
significantly generalize prior formulations, such as block MDPs, for representat
ion learning in RL. Algorithmically, we develop FLAMBE, which engages in explora
tion and representation learning for provably efficient RL in low rank transitio
n models. On a technical level, our analysis eliminates reachability assumptions
 that appear in prior results on the simpler block MDP model and may be of indep
endent interest.
**********

A Class of Algorithms for General Instrumental Variable Models
Niki Kilbertus, Matt J. Kusner, Ricardo Silva
Causal treatment effect estimation is a key problem that arises in a variety of
real-world settings, from personalized medicine to governmental policy making. T
here has been a flurry of recent work in machine learning on estimating causal e
ffects when one has access to an instrument. However, to achieve identifiability
, they in general require one-size-fits-all assumptions such as an additive erro
r model for the outcome. An alternative is partial identification, which provide
s bounds on the causal effect. Little exists in terms of bounding methods that c

an deal with the most general case, where the treatment itself can be continuous. Moreover, bounding methods generally do not allow for a continuum of assumptions on the shape of the causal effect that can smoothly trade off stronger background knowledge for more informative bounds. In this work, we provide a method for causal effect bounding in continuous distributions, leveraging recent advances in gradient-based methods for the optimization of computationally intractable objective functions. We demonstrate on a set of synthetic and real-world data that our bounds capture the causal effect when additive methods fail, providing a useful range of answers compatible with observation as opposed to relying on unwarranted structural assumptions.

**********

Black-Box Ripper: Copying black-box models using generative evolutionary algorithms

Antonio Barbalau, Adrian Cosma, Radu Tudor Ionescu, Marius Popescu

We study the task of replicating the functionality of black-box neural models, for which we only know the output class probabilities provided for a set of input images. We assume back-propagation through the black-box model is not possible and its training images are not available, e.g. the model could be exposed only through an API. In this context, we present a teacher-student framework that can distill the black-box (teacher) model into a student model with minimal accuracy loss. To generate useful data samples for training the student, our framework (i) learns to generate images on a proxy data set (with images and classes different from those used to train the black-box) and (ii) applies an evolutionary strategy to make sure that each generated data sample exhibits a high response for a specific class when given as input to the black box. Our framework is compared with several baseline and state-of-the-art methods on three benchmark data sets. The empirical evidence indicates that our model is superior to the considered baselines. Although our method does not back-propagate through the black-box network, it generally surpasses state-of-the-art methods that regard the teacher as a glass-box model. Our code is available at: https://github.com/antoniobarbalau/black-box-ripper.

**********

Bayesian Optimization of Risk Measures

Sait Cakmak, Raul Astudillo Marban, Peter Frazier, Enlu 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.

**********

TorsionNet: A Reinforcement Learning Approach to Sequential Conformer Search

Tarun Gogineni, Ziping Xu, Exequiel Punzalan, Runxuan Jiang, Joshua Kammeraad, Ambuj Tewari, Paul Zimmerman

Molecular geometry prediction of flexible molecules, or conformer search, is a long-standing challenge in computational chemistry. This task is of great importance for predicting structure-activity relationships for a wide variety of substances ranging from biomolecules to ubiquitous materials. Substantial computational resources are invested in Monte Carlo and Molecular Dynamics methods to generate diverse and representative conformer sets for medium to large molecules, which are yet intractable to chemoinformatic conformer search methods. We present TorsionNet, an efficient sequential conformer search technique based on reinforcement learning under the rigid rotor approximation. The model is trained via curriculum learning, whose theoretical benefit is explored in detail, to maximize a novel metric grounded in thermodynamics called the Gibbs Score. Our experimental results show that TorsionNet outperforms the highest-scoring chemoinformatics method by 4x on large branched alkanes, and by several orders of magnitude on the previously unexplored biopolymer lignin, with applications in renewable energy. TorsionNet also outperforms the far more exhaustive but computationally intensive Self-Guided Molecular Dynamics sampling method.

**********

GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis

Katja Schwarz, Yiyi Liao, Michael Niemeyer, Andreas Geiger
While 2D generative adversarial networks have enabled high-resolution image synthesis, they largely lack an understanding of the 3D world and the image formation process. Thus, they do not provide precise control over camera viewpoint or object pose. To address this problem, several recent approaches leverage intermediate voxel-based representations in combination with differentiable rendering. However, existing methods either produce low image resolution or fall short in disentangling camera and scene properties, e.g., the object identity may vary with the viewpoint. In this paper, we propose a generative model for radiance fields which have recently proven successful for novel view synthesis of a single scene. In contrast to voxel-based representations, radiance fields are not confined to a coarse discretization of the 3D space, yet allow for disentangling camera and scene properties while degrading gracefully in the presence of reconstruction ambiguity. By introducing a multi-scale patch-based discriminator, we demonstrate synthesis of high-resolution images while training our model from unposed 2D images alone. We systematically analyze our approach on several challenging synthetic and real-world datasets. Our experiments reveal that radiance fields are a powerful representation for generative image synthesis, leading to 3D consistent models that render with high fidelity.
**********

PIE-NET: Parametric Inference of Point Cloud Edges
Xiaogang Wang, Yuelang Xu, Kai Xu, Andrea Tagliasacchi, Bin Zhou, Ali Mahdavi-Amiri, Hao Zhang
We introduce an end-to-end learnable technique to robustly identify feature edges in 3D point cloud data. We represent these edges as a collection of parametric curves (i.e.,~lines, circles, and B-splines). Accordingly, our deep neural network, coined PIE-NET, is trained for parametric inference of edges. The network relies on a "region proposal" architecture, where a first module proposes an over-complete collection of edge and corner points, and a second module ranks each proposal to decide whether it should be considered. We train and evaluate our method on the ABC dataset, a large dataset of CAD models, and compare our results to those produced by traditional (non-learning) processing pipelines, as well as a recent deep learning based edge detector (EC-NET). Our results significantly improve over the state-of-the-art from both a quantitative and qualitative standpoint.
**********

A Simple Language Model for Task-Oriented Dialogue
Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, Richard Socher
Task-oriented dialogue is often decomposed into three tasks: understanding user input, deciding actions, and generating a response. While such decomposition might suggest a dedicated model for each sub-task, we find a simple, unified approach leads to state-of-the-art performance on the MultiWOZ dataset. SimpleTOD is a simple approach to task-oriented dialogue that uses a single, causal language model trained on all sub-tasks recast as a single sequence prediction problem. This allows SimpleTOD to fully leverage transfer learning from pre-trained, open domain, causal language models such as GPT-2. SimpleTOD improves over the prior state-of-the-art in joint goal accuracy for dialogue state tracking, and our analysis reveals robustness to noisy annotations in this setting. SimpleTOD also improves the main metrics used to evaluate action decisions and response generation in an end-to-end setting: inform rate by 8.1 points, success rate by 9.7 points, and combined score by 7.2 points.
**********

A Continuous-Time Mirror Descent Approach to Sparse Phase Retrieval
Fan Wu, 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.
**********

Confidence sequences for sampling without replacement

Ian Waudby-Smith, Aaditya Ramdas
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 mean-field analysis of two-player zero-sum games
Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, Joan Bruna
Finding Nash equilibria in two-player zero-sum continuous games is a central problem in machine learning, e.g. for training both GANs and robust models. The existence of pure Nash equilibria requires strong conditions which are not typically met in practice. Mixed Nash equilibria exist in greater generality and may be found using mirror descent. Yet this approach does not scale to high dimensions. To address this limitation, we parametrize mixed strategies as mixtures of particles, whose positions and weights are updated using gradient descent-ascent. We study this dynamics as an interacting gradient flow over measure spaces endowed with the Wasserstein-Fisher-Rao metric. We establish global convergence to an approximate equilibrium for the related Langevin gradient-ascent dynamic. We prove a law of large numbers that relates particle dynamics to mean-field dynamics. Our method identifies mixed equilibria in high dimensions and is demonstrably effective for training mixtures of GANs.
**********

Leap-Of-Thought: Teaching Pre-Trained Models to Systematically Reason Over Implicit Knowledge
Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, Jonathan Berant
To what extent can a neural network systematically reason over symbolic facts? Evidence suggests that large pre-trained language models (LMs) acquire some reasoning capacity, but this ability is difficult to control.
Recently, it has been shown that Transformer-based models succeed in consistent reasoning over explicit symbolic facts, under a "closed-world" assumption.
However, in an open-domain setup, it is desirable to tap into the vast reservoir of implicit knowledge already encoded in the parameters of pre-trained LMs.
In this work, we provide a first demonstration that LMs can be trained to reliably perform systematic reasoning combining both implicit, pre-trained knowledge and explicit natural language statements.
To do this, we describe a procedure for automatically generating datasets that teach a model new reasoning skills, and demonstrate that models learn to effectively perform inference which involves implicit taxonomic and world knowledge, chaining and counting.
Finally, we show that "teaching" models to reason generalizes beyond the training distribution: they successfully compose the usage of multiple reasoning skills in single examples.
Our work paves a path towards open-domain systems that constantly improve by interacting with users who can instantly correct a model by adding simple natural language statements.
**********

Pipeline PSRO: A Scalable Approach for Finding Approximate Nash Equilibria in Large Games
Stephen Mcaleer, JB Lanier, Roy Fox, Pierre Baldi
Finding approximate Nash equilibria in zero-sum imperfect-information games is challenging when the number of information states is large. Policy Space Response Oracles (PSRO) is a deep reinforcement learning algorithm grounded in game theory that is guaranteed to converge to an approximate Nash equilibrium. However, PSRO requires training a reinforcement learning policy at each iteration, making it too slow for large games. We show through counterexamples and experiments that DCH and Rectified PSRO, two existing approaches to scaling up PSRO, fail to converge even in small games. We introduce Pipeline PSRO (P2SRO), the first scalable PSRO-based method for finding approximate Nash equilibria in large zero-sum imperfect-information games. P2SRO is able to parallelize PSRO with convergence guarantees by maintaining a hierarchical pipeline of reinforcement learning worke

rs, each training against the policies generated by lower levels in the hierarch
y. We show that unlike existing methods, P2SRO converges to an approximate Nash
equilibrium, and does so faster as the number of parallel workers increases, acr
oss a variety of imperfect information games. We also introduce an open-source e
nvironment for Barrage Stratego, a variant of Stratego with an approximate game
tree complexity of 10^50. P2SRO is able to achieve state-of-the-art performance
on Barrage Stratego and beats all existing bots. Experiment code is available at
 https://github.com/JBLanier/pipeline-psro.
**********

Improving Sparse Vector Technique with Renyi Differential Privacy
Yuqing Zhu, Yu-Xiang Wang
The Sparse Vector Technique (SVT) is one of the most fundamental algorithmic too
ls in differential privacy (DP). It also plays a central role in the state-of-th
e-art algorithms for adaptive data analysis and model-agnostic private learning.
 In this paper, we revisit SVT from the lens of Renyi differential privacy, whic
h results in new privacy bounds, new theoretical insight and new variants of SVT
 algorithms. A notable example is a Gaussian mechanism version of SVT, which pro
vides better utility over the standard (Laplace-mechanism-based) version thanks
to its more concentrated noise and tighter composition. Extensive empirical eval
uation demonstrates the merits of Gaussian SVT over the Laplace SVT and other al
ternatives, which encouragingly suggests that using Gaussian SVT as a drop-in re
placement could make SVT-based algorithms practical in downstream tasks.
**********

Latent Template Induction with Gumbel-CRFs
Yao Fu, Chuanqi Tan, Bin Bi, Mosha Chen, Yansong Feng, Alexander Rush
Learning to control the structure of sentences is a challenging problem in text
generation. Existing work either relies on simple deterministic approaches or RL
-based hard structures. We explore the use of structured variational autoencoder
s to infer latent templates for sentence generation using a soft, continuous rel
axation in order to utilize reparameterization for training. Specifically, we pr
opose a Gumbel-CRF, a continuous relaxation of the CRF sampling algorithm using
a relaxed Forward-Filtering Backward-Sampling (FFBS) approach. As a reparameteri
zed gradient estimator, the Gumbel-CRF gives more stable gradients than score-fu
nction based estimators. As a structured inference network, we show that it lear
ns interpretable templates during training, which allows us to control the decod
er during testing. We demonstrate the effectiveness of our methods with experime
nts on data-to-text generation and unsupervised paraphrase generation.
**********

Instance Based Approximations to Profile Maximum Likelihood
Nima Anari, Moses Charikar, Kirankumar Shiragur, Aaron Sidford
In this paper we provide a new efficient algorithm for approximately computing t
he profile maximum likelihood (PML) distribution, a prominent quantity in symmet
ric property estimation. We provide an algorithm which matches the previous best
 known efficient algorithms for computing approximate PML distributions and impr
oves when the number of distinct observed frequencies in the given instance is s
mall. We achieve this result by exploiting new sparsity structure in approximate
 PML distributions and providing a new matrix rounding algorithm, of independent
 interest. Leveraging this result, we obtain the first provable computationally
efficient implementation of PseudoPML, a general framework for estimating a broa
d class of symmetric properties. Additionally, we obtain efficient PML-based est
imators for distributions with small profile entropy, a natural instance-based c
omplexity measure. Further, we provide a simpler and more practical PseudoPML im
plementation that matches the best-known theoretical guarantees of such an estim
ator and evaluate this method empirically.
**********

Factorizable Graph Convolutional Networks
Yiding Yang, Zunlei Feng, Mingli Song, Xinchao Wang
Graphs have been widely adopted to denote structural connections between entitie
s. The relations are in many cases heterogeneous, but entangled together and den
oted merely as a single edge between a pair of nodes. For example, in a social n

etwork graph, users in different latent relationships like friends and colleague
s, are usually connected via a bare edge that conceals such intrinsic connection
s. In this paper, we introduce a novel graph convolutional network (GCN), termed
 as factorizable graph convolutional network (FactorGCN), that explicitly disent
angles such intertwined relations encoded in a graph. FactorGCN takes a simple g
raph as input, and disentangles it into several factorized graphs, each of which
 represents a latent and disentangled relation among nodes. The features of the
nodes are then aggregated separately in each factorized latent space to produce
disentangled features, which further leads to better performances for downstream
 tasks. We evaluate the proposed FactorGCN both qualitatively and quantitatively
 on the synthetic and real-world datasets, and demonstrate that it yields truly
encouraging results in terms of both disentangling and feature aggregation. Code
 is publicly available at https://github.com/ihollywhy/FactorGCN.PyTorch.
**********

Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses
Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, Venkatesh Babu R
Advances in the development of adversarial attacks have been fundamental to the
progress of adversarial defense research. Efficient and effective attacks are cr
ucial for reliable evaluation of defenses, and also for developing robust models
. Adversarial attacks are often generated by maximizing standard losses such as
the cross-entropy loss or maximum-margin loss within a constraint set using Proj
ected Gradient Descent (PGD). In this work, we introduce a relaxation term to th
e standard loss, that finds more suitable gradient-directions, increases attack
efficacy and leads to more efficient adversarial training. We propose Guided Adv
ersarial Margin Attack (GAMA), which utilizes function mapping of the clean imag
e to guide the generation of adversaries, thereby resulting in stronger attacks.
 We evaluate our attack against multiple defenses and show improved performance
when compared to existing attacks. Further, we propose Guided Adversarial Traini
ng (GAT), which achieves state-of-the-art performance amongst single-step defens
es by utilizing the proposed relaxation term for both attack generation and trai
ning.
**********

A Study on Encodings for Neural Architecture Search
Colin White, Willie Neiswanger, Sam Nolen, Yash Savani
Neural architecture search (NAS) has been extensively studied in the past few ye
ars. A popular approach is to represent each neural architecture in the search s
pace as a directed acyclic graph (DAG), and then search over all DAGs by encodin
g the adjacency matrix and list of operations as a set of hyperparameters. Recen
t work has demonstrated that even small changes to the way each architecture is
encoded can have a significant effect on the performance of NAS algorithms.
**********

Noise2Same: Optimizing A Self-Supervised Bound for Image Denoising
Yaochen Xie, Zhengyang Wang, Shuiwang Ji
Self-supervised frameworks that learn denoising models with merely individual no
isy images have shown strong capability and promising performance in various ima
ge denoising tasks. Existing self-supervised denoising frameworks are mostly bui
lt upon the same theoretical foundation, where the denoising models are required
 to be J-invariant. However, our analyses indicate that the current theory and t
he J-invariance may lead to denoising models with reduced performance. In this w
ork, we introduce Noise2Same, a novel self-supervised denoising framework. In No
ise2Same, a new self-supervised loss is proposed by deriving a self-supervised u
pper bound of the typical supervised loss. In particular, Noise2Same requires ne
ither J-invariance nor extra information about the noise model and can be used i
n a wider range of denoising applications. We analyze our proposed Noise2Same bo
th theoretically and experimentally. The experimental results show that our Nois
e2Same remarkably outperforms previous self-supervised denoising methods in term
s of denoising performance and training efficiency.
**********

Early-Learning Regularization Prevents Memorization of Noisy Labels
Sheng Liu, Jonathan Niles-Weed, Narges Razavian, Carlos Fernandez-Granda

We propose a novel framework to perform classification via deep learning in the presence of noisy annotations. When trained on noisy labels, deep neural networks have been observed to first fit the training data with clean labels during an "early learning" phase, before eventually memorizing the examples with false labels. We prove that early learning and memorization are fundamental phenomena in high-dimensional classification tasks, even in simple linear models, and give a theoretical explanation in this setting. Motivated by these findings, we develop a new technique for noisy classification tasks, which exploits the progress of the early learning phase. In contrast with existing approaches, which use the model output during early learning to detect the examples with clean labels, and either ignore or attempt to correct the false labels, we take a different route and instead capitalize on early learning via regularization. There are two key elements to our approach. First, we leverage semi-supervised learning techniques to produce target probabilities based on the model outputs. Second, we design a regularization term that steers the model towards these targets, implicitly preventing memorization of the false labels. The resulting framework is shown to provide robustness to noisy annotations on several standard benchmarks and real-world datasets, where it achieves results comparable to the state of the art.
**********

LAPAR: Linearly-Assembled Pixel-Adaptive Regression Network for Single Image Super-resolution and Beyond
Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, Jiaya Jia
Single image super-resolution (SISR) deals with a fundamental problem of upsampling a low-resolution (LR) image to its high-resolution (HR) version. Last few years have witnessed impressive progress propelled by deep learning methods. However, one critical challenge faced by existing methods is to strike a sweet spot of deep model complexity and resulting SISR quality. This paper addresses this pain point by proposing a linearly-assembled pixel-adaptive regression network (LAPAR), which casts the direct LR to HR mapping learning into a linear coefficient regression task over a dictionary of multiple predefined filter bases. Such a parametric representation renders our model highly lightweight and easy to optimize while achieving state-of-the-art results on SISR benchmarks. Moreover, based on the same idea, LAPAR is extended to tackle other restoration tasks, e.g., image denoising and JPEG image deblocking, and again, yields strong performance.
**********

Learning Parities with Neural Networks
Amit Daniely, Eran Malach
In recent years we see a rapidly growing line of research which shows learnability of various models via common neural network algorithms. Yet, besides a very few outliers, these results show learnability of models that can be learned using linear methods. Namely, such results show that learning neural-networks with gradient-descent is competitive with learning a linear classifier on top of a data-independent representation of the examples. This leaves much to be desired, as neural networks are far more successful than linear methods. Furthermore, on the more conceptual level, linear models don't seem to capture the``deepness" of deep networks. In this paper we make a step towards showing leanability of models that are inherently non-linear. We show that under certain distributions, sparse parities are learnable via gradient decent on depth-two network. On the other hand, under the same distributions, these parities cannot be learned efficiently by linear methods.
**********

Consistent Plug-in Classifiers for Complex Objectives and Constraints
Shiv Kumar Tavker, Harish Guruprasad Ramaswamy, Harikrishna Narasimhan
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Movement Pruning: Adaptive Sparsity by Fine-Tuning
Victor Sanh, Thomas Wolf, Alexander Rush

Magnitude pruning is a widely used strategy for reducing model size in pure supe
rvised learning; however, it is less effective in the transfer learning regime t
hat has become standard for state-of-the-art natural language processing applica
tions. We propose the use of movement pruning, a simple, deterministic first-ord
er weight pruning method that is more adaptive to pretrained model fine-tuning.
We give mathematical foundations to the method and compare it to existing zeroth
- and first-order pruning methods. Experiments show that when pruning large pret
rained language models, movement pruning shows significant improvements in high-
sparsity regimes. When combined with distillation, the approach achieves minimal
 accuracy loss with down to only 3% of the model parameters.
**********

Sanity-Checking Pruning Methods: Random Tickets can Win the Jackpot
Jingtong Su, Yihang Chen, Tianle Cai, Tianhao Wu, Ruiqi Gao, Liwei Wang, Jason D
. Lee
Network pruning is a method for reducing test-time computational resource requir
ements with minimal performance degradation. Conventional wisdom of pruning algo
rithms suggests that: (1) Pruning methods exploit information from training data
 to find good subnetworks; (2) The architecture of the pruned network is crucial
 for good performance. In this paper, we conduct sanity checks for the above bel
iefs on several recent unstructured pruning methods and surprisingly find that:
(1) A set of methods which aims to find good subnetworks of the randomly-initial
ized network (which we call initial tickets''), hardly exploits any information
from the training data; (2) For the pruned networks obtained by these methods, r
andomly changing the preserved weights in each layer, while keeping the total nu
mber of preserved weights unchanged per layer, does not affect the final perform
ance. These findings inspire us to choose a series of simple \emph{data-independ
ent} prune ratios for each layer, and randomly prune each layer accordingly to g
et a subnetwork (which we callrandom tickets''). Experimental results show that
our zero-shot random tickets outperforms or attains similar performance compared
 to existing initial tickets''. In addition, we identify one existing pruning me
thod that passes our sanity checks. We hybridize the ratios in our random ticket
 with this method and propose a new method calledhybrid tickets'', which achieve
s further improvement.
**********

Online Matrix Completion with Side Information
Mark Herbster, Stephen Pasteris, Lisa Tse
We give an online algorithm and prove novel mistake and regret bounds for online
 binary matrix completion with side information. The mistake bounds we prove are
 of the form \tilde{O}(D/\gamma^2). The term 1/\gamma^2 is analogous to the usua
l margin term in SVM (perceptron) bounds. More specifically, if we assume that t
here is some factorization of the underlying m x n matrix into PQ^T, where the r
ows of P are interpreted as "classifiers" in R^d and the rows of Q as "instances
" in R^d, then gamma is the maximum (normalized) margin over all factorizations
PQ^T consistent with the observed matrix. The quasi-dimension term D measures th
e quality of side information. In the presence of vacuous side information, D =
m+n. However, if the side information is predictive of the underlying factorizat
ion of the matrix, then in an ideal case, D \in O(k + l) where k is the number o
f distinct row factors and l is the number of distinct column factors. We additi
onally provide a generalization of our algorithm to the inductive setting. In th
is setting, we provide an example where the side information is not directly spe
cified in advance. For this example, the quasi-dimension D is now bounded by O(k
^2 + l^2).
**********

Position-based Scaled Gradient for Model Quantization and Pruning
Jangho Kim, KiYoon Yoo, Nojun Kwak
We propose the position-based scaled gradient (PSG) that scales the gradient dep
ending on the position of a weight vector to make it more compression-friendly.
First, we theoretically show that applying PSG to the standard gradient descent
(GD), which is called PSGD, is equivalent to the GD in the warped weight space,
a space made by warping the original weight space via an appropriately designed

invertible function. Second, we empirically show that PSG acting as a regularize
r to a weight vector is favorable for model compression domains such as quantiza
tion and pruning. PSG reduces the gap between the weight distributions of a full
-precision model and its compressed counterpart. This enables the versatile depl
oyment of a model either as an uncompressed mode or as a compressed mode dependi
ng on the availability of resources. The experimental results on CIFAR-10/100 an
d ImageNet datasets show the effectiveness of the proposed PSG in both domains o
f pruning and quantization even for extremely low bits. The code is released in
Github.
**********

Online Learning with Primary and Secondary Losses
Avrim Blum, Han Shao
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Graph Information Bottleneck
Tailin Wu, Hongyu Ren, Pan Li, Jure Leskovec
Representation learning of graph-structured data is challenging because both gra
ph structure and node features carry important information. Graph Neural Network
s (GNNs) provide an expressive way to fuse information from network structure an
d node features. However, GNNs are prone to adversarial attacks. Here we introdu
ce Graph Information Bottleneck (GIB), an information-theoretic principle that o
ptimally balances expressiveness and robustness of the learned representation of
 graph-structured data. Inheriting from the general Information Bottleneck (IB),
 GIB aims to learn the minimal sufficient representation for a given task by max
imizing the mutual information between the representation and the target, and si
multaneously constraining the mutual information between the representation and
the input data. Different from the general IB, GIB regularizes the structural as
 well as the feature information. We design two sampling algorithms for structur
al regularization and instantiate the GIB principle with two new models: GIB-Cat
 and GIB-Bern, and demonstrate the benefits by evaluating the resilience to adve
rsarial attacks. We show that our proposed models are more robust than state-of-
the-art graph defense models. GIB-based models empirically achieve up to 31% imp
rovement with adversarial perturbation of the graph structure as well as node fe
atures.
**********

The Complexity of Adversarially Robust Proper Learning of Halfspaces with Agnost
ic Noise
Ilias Diakonikolas, Daniel M. 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Adaptive Online Estimation of Piecewise Polynomial Trends
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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

RNNPool: Efficient Non-linear Pooling for RAM Constrained Inference
Oindrila Saha, Aditya Kusupati, Harsha Vardhan Simhadri, Manik Varma, Prateek Ja
in
Standard Convolutional Neural Networks (CNNs) designed for computer vision tasks
 tend to have large intermediate activation maps. These require large working me
mory and are thus unsuitable for deployment on resource-constrained devices typi
cally used for inference on the edge. Aggressively downsampling the images via p

ooling or strided convolutions can address the problem but leads to a significan t decrease in accuracy due to gross aggregation of the feature map by standard p ooling operators. In this paper, we introduce RNNPool, a novel pooling operator based on Recurrent Neural Networks (RNNs), that efficiently aggregates features over large patches of an image and rapidly downsamples activation maps. Empiric al evaluation indicates that an RNNPool layer can effectively replace multiple b locks in a variety of architectures such as MobileNets, DenseNet when applied to standard vision tasks like image classification and face detection. That is, RN NPool can significantly decrease computational complexity and peak memory usage for inference while retaining comparable accuracy. We use RNNPool with the stand ard S3FD architecture to construct a face detection method that achieves state-o f-the-art MAP for tiny ARM Cortex-M4 class microcontrollers with under 256 KB of RAM. Code is released at https://github.com/Microsoft/EdgeML.
**********

Agnostic Learning with Multiple Objectives
Corinna Cortes, Mehryar Mohri, Javier Gonzalvo, Dmitry Storcheus
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

3D Multi-bodies: Fitting Sets of Plausible 3D Human Models to Ambiguous Image Da ta
Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, Andr ea Vedaldi
We consider the problem of obtaining dense 3D reconstructions of deformable obje cts from single and partially occluded views. In such cases, the visual evidence is usually insufficient to identify a 3D reconstruction uniquely, so we aim at recovering several plausible reconstructions compatible with the input data. We suggest that ambiguities can be modeled more effectively by parametrizing the po ssible body shapes and poses via a suitable 3D model, such as SMPL for humans. We propose to learn a multi-hypothesis neural network regressor using a best-of-M loss, where each of the M hypotheses is constrained to lie on a manifold of pl ausible human poses by means of a generative model. We show that our method outp erforms alternative approaches in ambiguous pose recovery on standard benchmarks for 3D humans, and in heavily occluded versions of these benchmarks.
**********

Auto-Panoptic: Cooperative Multi-Component Architecture Search for Panoptic Segm entation
Yangxin Wu, Gengwei Zhang, Hang Xu, Xiaodan Liang, Liang Lin
Panoptic segmentation is posed as a new popular test-bed for the state-of-the-ar t holistic scene understanding methods with the requirement of simultaneously se gmenting both foreground things and background stuff. The state-of-the-art panop tic segmentation network exhibits high structural complexity in different networ k components, i.e. backbone, proposal-based foreground branch, segmentation-base d background branch, and feature fusion module across branches, which heavily re lies on expert knowledge and tedious trials. In this work, we propose an efficie nt, cooperative and highly automated framework to simultaneously search for all main components including backbone, segmentation branches, and feature fusion mo dule in a unified panoptic segmentation pipeline based on the prevailing one-sho t Network Architecture Search (NAS) paradigm. Notably, we extend the common sing le-task NAS into the multi-component scenario by taking the advantages of the ne wly proposed intra-modular search space and problem-oriented inter-modular searc h space, which helps us to obtain an optimal network architecture that not only performs well in both instance segmentation and semantic segmentation tasks but also be aware of the reciprocal relations between foreground things and backgrou nd stuff classes. To relieve the vast computation burden incurred by applying NA S to complicated network architectures, we present a novel path-priority greedy search policy to find a robust, transferrable architecture with significantly re duced searching overhead. Our searched architecture, namely Auto-Panoptic, achie

ves the new state-of-the-art on the challenging COCO and ADE20K benchmarks. More over, extensive experiments are conducted to demonstrate the effectiveness of pa th-priority policy and transferability of Auto-Panoptic across different dataset s.
**********
Differentiable Top-k with Optimal Transport
Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, To mas Pfister
Finding the k largest or smallest elements from a collection of scores, i.e., to p-k operation, is an important model component widely used in information retrie val, machine learning, and data mining. However, if the top-k operation is imple mented in an algorithmic way, e.g., using bubble algorithm, the resulted model c annot be trained in an end-to-end way using prevalent gradient descent algorithm s. This is because these implementations typically involve swapping indices, who se gradient cannot be computed. Moreover, the corresponding mapping from the inp ut scores to the indicator vector of whether this element belongs to the top-k s et is essentially discontinuous. To address the issue, we propose a smoothed app roximation, namely SOFT (Scalable Optimal transport-based diFferenTiable) top-k operator. Specifically, our SOFT top-k operator approximates the output of top-k operation as the solution of an Entropic Optimal Transport (EOT) problem. The g radient of the SOFT operator can then be efficiently approximated based on the o ptimality conditions of EOT problem.
We then apply the proposed operator to k-nearest neighbors algorithm and beam se arch algorithm. The numerical experiment demonstrates their achieve improved per formance.
**********
Information-theoretic Task Selection for Meta-Reinforcement Learning
Ricardo Luna Gutierrez, Matteo Leonetti
In Meta-Reinforcement Learning (meta-RL) an agent is trained on a set of tasks t o prepare for and learn faster in new, unseen, but related tasks. The training t asks are usually hand-crafted to be representative of the expected distribution of target tasks and hence all used in training. We show that given a set of trai ning tasks, learning can be both faster and more effective (leading to better pe rformance in the target tasks), if the training tasks are appropriately selected . We propose a task selection algorithm based on information theory, which optim izes the set of tasks used for training in meta-RL, irrespectively of how they a re generated. The algorithm establishes which training tasks are both sufficient ly relevant for the target tasks, and different enough from one another. We repr oduce different meta-RL experiments from the literature and show that our task s election algorithm improves the final performance in all of them.
**********
A Limitation of the PAC-Bayes Framework
Roi Livni, 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-auth ors prior to requesting a name change in the electronic proceedings.
**********
On Completeness-aware Concept-Based Explanations in Deep Neural Networks
Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, Pradeep Ravi kumar
Human explanations of high-level decisions are often expressed in terms of key c oncepts the decisions are based on. In this paper, we study such concept-based e xplainability for Deep Neural Networks (DNNs). First, we define the notion of \e mph{completeness}, which quantifies how sufficient a particular set of concepts is in explaining a model's prediction behavior based on the assumption that comp lete concept scores are sufficient statistics of the model prediction. Next, we propose a concept discovery method that aims to infer a complete set of concepts that are additionally encouraged to be interpretable, which addresses the limit ations of existing methods on concept explanations. To define an importance scor

e for each discovered concept, we adapt game-theoretic notions to aggregate over sets and propose \emph{ConceptSHAP}. Via proposed metrics and user studies, on a synthetic dataset with apriori-known concept explanations, as well as on real-world image and language datasets, we validate the effectiveness of our method in finding concepts that are both complete in explaining the decisions and interpretable.

**********

Stochastic Recursive Gradient Descent Ascent for Stochastic Nonconvex-Strongly-Concave Minimax Problems

Luo Luo, Haishan Ye, Zhichao Huang, Tong Zhang

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

**********

Why Normalizing Flows Fail to Detect Out-of-Distribution Data

Polina Kirichenko, Pavel Izmailov, Andrew G. Wilson

Detecting out-of-distribution (OOD) data is crucial for robust machine learning systems. Normalizing flows are flexible deep generative models that often surprisingly fail to distinguish between in- and out-of-distribution data: a flow trained on pictures of clothing assigns higher likelihood to handwritten digits. We investigate why normalizing flows perform poorly for OOD detection. We demonstrate that flows learn local pixel correlations and generic image-to-latent-space transformations which are not specific to the target image datasets, focusing on flows based on coupling layers. We show that by modifying the architecture of flow coupling layers we can bias the flow towards learning the semantic structure of the target data, improving OOD detection. Our investigation reveals that properties that enable flows to generate high-fidelity images can have a detrimental effect on OOD detection.

**********

Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay

Joao Marques-Silva, Thomas Gerspacher, Martin Cooper, Alexey Ignatiev, Nina Narodytska

Recent work proposed the computation of so-called PI-explanations of Naive Bayes Classifiers (NBCs). PI-explanations are subset-minimal sets of feature-value pairs that are sufficient for the prediction, and have been computed with state-of-the-art exact algorithms that are worst-case exponential in time and space. In contrast, we show that the computation of one PI-explanation for an NBC can be achieved in log-linear time, and that the same result also applies to the more general class of linear classifiers. Furthermore, we show that the enumeration of PI-explanations can be obtained with polynomial delay. Experimental results demonstrate the performance gains of the new algorithms when compared with earlier work. The experimental results also investigate ways to measure the quality of heuristic explanations.

**********

Unsupervised Translation of Programming Languages

Baptiste Roziere, Marie-Anne Lachaux, Lowik Chanussot, Guillaume Lample

A transcompiler, also known as source-to-source translator, is a system that converts source code from a high-level programming language (such as C++ or Python) to another. Transcompilers are primarily used for interoperability, and to port codebases written in an obsolete or deprecated language (e.g. COBOL, Python 2) to a modern one. They typically rely on handcrafted rewrite rules, applied to the source code abstract syntax tree. Unfortunately, the resulting translations often lack readability, fail to respect the target language conventions, and require manual modifications in order to work properly. The overall translation process is time-consuming and requires expertise in both the source and target languages, making code-translation projects expensive. Although neural models significantly outperform their rule-based counterparts in the context of natural language translation, their applications to transcompilation have been limited due to t

he scarcity of parallel data in this domain. In this paper, we propose to levera ge recent approaches in unsupervised machine translation to train a fully unsupe rvised neural transcompiler. We train our model on source code from open source GitHub projects, and show that it can translate functions between C++, Java, and Python with high accuracy. Our method relies exclusively on monolingual source code, requires no expertise in the source or target languages, and can easily be generalized to other programming languages. We also build and release a test se t composed of 852 parallel functions, along with unit tests to check the correct ness of translations. We show that our model outperforms rule-based commercial b aselines by a significant margin.
**********

Adversarial Style Mining for One-Shot Unsupervised Domain Adaptation
Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, Yi Yang
We aim at the problem named One-Shot Unsupervised Domain Adaptation. Unlike trad itional Unsupervised Domain Adaptation, it assumes that only one unlabeled targe t sample can be available when learning to adapt. This setting is realistic but more challenging, in which conventional adaptation approaches are prone to failu re due to the scarce of unlabeled target data. To this end, we propose a novel A dversarial Style Mining approach, which combines the style transfer module and t ask-specific module into an adversarial manner. Specifically, the style transfer module iteratively searches for harder stylized images around the one-shot targ et sample according to the current learning state, leading the task model to exp lore the potential styles that are difficult to solve in the almost unseen targe t domain,
thus boosting the adaptation performance in a data-scarce scenario. The adversar ial learning framework makes the style transfer module and task-specific module benefit each other during the competition. Extensive experiments on both cross-d omain classification and segmentation benchmarks verify that ASM achieves state- of-the-art adaptation performance under the challenging one-shot setting.
**********

Optimally Deceiving a Learning Leader in Stackelberg Games
Georgios Birmpas, Jiarui Gan, Alexandros Hollender, Francisco Marmolejo, Ninad R ajgopal, Alexandros Voudouris
Recent results in the ML community have revealed that learning algorithms used t o compute the optimal strategy for the leader to commit to in a Stackelberg game , are susceptible to manipulation by the follower. Such a learning algorithm ope rates by querying the best responses or the payoffs of the follower, who consequ ently can deceive the algorithm by responding as if their payoffs were much diff erent than what they actually are. For this strategic behavior to be successful, the main challenge faced by the follower is to pinpoint the payoffs that would make the learning algorithm compute a commitment so that best responding to it m aximizes the follower's utility, according to the true payoffs. While this probl em has been considered before, the related literature only focused on the simpli fied scenario in which the payoff space is finite, thus leaving the general vers ion of the problem unanswered. In this paper, we fill this gap by showing that i t is always possible for the follower to efficiently compute (near-)optimal payo ffs for various scenarios of learning interaction between the leader and the fol lower.
**********

Online Optimization with Memory and Competitive Control
Guanya Shi, Yiheng Lin, Soon-Jo Chung, Yisong Yue, 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.
**********

IDEAL: Inexact DEcentralized Accelerated Augmented Lagrangian Method
Yossi Arjevani, Joan Bruna, Bugra Can, Mert Gurbuzbalaban, Stefanie Jegelka, Hon gzhou Lin
We introduce a framework for designing primal methods under the decentralized op

timization setting where local functions are smooth and strongly convex. Our app
roach consists of approximately solving a sequence of sub-problems induced by th
e accelerated augmented Lagrangian method, thereby providing a systematic way fo
r deriving several well-known decentralized algorithms including EXTRA and SSDA.
 When coupled with accelerated gradient descent, our framework yields a novel pr
imal algorithm whose convergence rate is optimal and matched by recently derived
 lower bounds. We provide experimental results that demonstrate the effectivenes
s of the proposed algorithm on highly ill-conditioned problems.
**********

Evolving Graphical Planner: Contextual Global Planning for Vision-and-Language N
avigation

Zhiwei Deng, Karthik Narasimhan, Olga Russakovsky

The ability to perform effective planning is crucial for building an instruction
-following agent. When navigating through a new environment, an agent is challen
ged with (1) connecting the natural language instructions with its progressively
 growing knowledge of the world; and (2) performing long-range planning and deci
sion making in the form of effective exploration and error correction. Current m
ethods are still limited on both fronts despite extensive efforts.  In this pape
r, we introduce Evolving Graphical Planner (EGP), a module that allows global pl
anning for navigation based on raw sensory input. The module dynamically constru
cts a graphical representation, generalizes the local action space to allow for
more flexible decision making, and performs efficient planning on a proxy repres
entation. We demonstrate our model on a challenging Vision-and-Language Navigati
on (VLN) task with photorealistic images, and achieve superior performance compa
red to previous navigation architectures. Concretely, we achieve 53% success rat
e on the test split of Room-to-Room navigation task (Anderson et al.) through pu
re imitation learning, outperforming previous architectures by up to 5%.
**********

Learning from Failure: De-biasing Classifier from Biased Classifier

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, Jinwoo Shin

Neural networks often learn to make predictions that overly rely on spurious cor
re- lation existing in the dataset, which causes the model to be biased. While p
revious work tackles this issue by using explicit labeling on the spuriously cor
related attributes or presuming a particular bias type, we instead utilize a che
aper, yet generic form of human knowledge, which can be widely applicable to var
ious types of bias. We first observe that neural networks learn to rely on the s
purious correlation only when it is “easier” to learn than the desired knowledge
, and such reliance is most prominent during the early phase of training. Based
on the obser- vations, we propose a failure-based debiasing scheme by training a
 pair of neural networks simultaneously. Our main idea is twofold; (a) we intent
ionally train the first network to be biased by repeatedly amplifying its “preju
dice”, and (b) we debias the training of the second network by focusing on sampl
es that go against the prejudice of the biased network in (a). Extensive experim
ents demonstrate that our method significantly improves the training of network
against various types of biases in both synthetic and real-world datasets. Surpr
isingly, our framework even occasionally outperforms the debiasing methods requi
ring explicit supervision of the spuriously correlated attributes.
**********

Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-e
ncoder

Zhisheng Xiao, Qing Yan, Yali Amit

Deep probabilistic generative models enable modeling the likelihoods of very hig
h dimensional data. An important application of generative modeling should be th
e ability to detect out-of-distribution (OOD) samples by setting a threshold on
the likelihood. However, a recent study shows that probabilistic generative mode
ls can, in some cases, assign higher likelihoods on certain types of OOD samples
, making the OOD detection rules based on likelihood threshold problematic. To a
ddress this issue, several OOD detection methods have been proposed for deep gen
erative models. In this paper, we make the observation that some of these method
s fail when applied to generative models based on  Variational Auto-encoders (VA

E). As an alternative, we propose Likelihood Regret, an efficient OOD score for VAEs. We benchmark our proposed method over existing approaches, and empirical results suggest that our method obtains the best overall OOD detection performances compared with other OOD method applied on VAE.
**********

Deep Diffusion-Invariant Wasserstein Distributional Classification
Sung Woo Park, Dong Wook Shu, Junseok Kwon
In this paper, we present a novel classification method called deep diffusion-invariant Wasserstein distributional classification (DeepWDC). DeepWDC represents input data and labels as probability measures to address severe perturbations in input data. It can output the optimal label measure in terms of diffusion invariance, where the label measure is stationary over time and becomes equivalent to a Gaussian measure. Furthermore, DeepWDC minimizes the 2-Wasserstein distance between the optimal label measure and Gaussian measure, which reduces the Wasserstein uncertainty. Experimental results demonstrate that DeepWDC can substantially enhance the accuracy of several baseline deterministic classification methods and outperforms state-of-the-art-methods on 2D and 3D data containing various types of perturbations (e.g., rotations, impulse noise, and down-scaling).
**********

Finding All $\epsilon$-Good Arms in Stochastic Bandits
Blake Mason, Lalit Jain, Ardhendu Tripathy, Robert Nowak
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Meta-Learning through Hebbian Plasticity in Random Networks
Elias Najarro, Sebastian Risi
Lifelong learning and adaptability are two defining aspects of biological agents. Modern reinforcement learning (RL) approaches
have shown significant progress in solving complex tasks, however once training is concluded, the found solutions are typically static and incapable of adapting to new information or perturbations. While it is still not completely understood how biological brains learn and adapt so efficiently from experience, it is believed that synaptic plasticity plays a prominent role in this process. Inspired by this biological mechanism, we propose a search method that, instead of optimizing the weight parameters of neural networks directly, only searches for synapse-specific Hebbian learning rules that allow the network to continuously self-organize its weights during the lifetime of the agent. We demonstrate our approach on several reinforcement learning tasks with different sensory modalities and more than 450K trainable plasticity parameters. We find that starting from completely random weights, the discovered Hebbian rules enable an agent to navigate a dynamical 2D-pixel environment; likewise they allow a simulated 3D quadrupedal robot to learn how to walk while adapting to morphological damage not seen during training and in the absence of any explicit reward or error signal in less than 100 timesteps.
**********

A Computational Separation between Private Learning and Online Learning
Mark Bun
A recent line of work has shown a qualitative equivalence between differentially private PAC learning and online learning: A concept class is privately learnable if and only if it is online learnable with a finite mistake bound. However, both directions of this equivalence incur significant losses in both sample and computational efficiency.
Studying a special case of this connection, Gonen, Hazan, and Moran (NeurIPS 2019) showed that uniform or highly sample-efficient pure-private learners can be time-efficiently compiled into online learners. We show that, assuming the existence of one-way functions, such an efficient conversion is impossible even for general pure-private learners with polynomial sample complexity. This resolves a question of Neel, Roth, and Wu (FOCS 2019).

**********
Top-KAST: Top-K Always Sparse Training
Siddhant Jayakumar, Razvan Pascanu, Jack Rae, Simon Osindero, Erich Elsen
Sparse neural networks are becoming increasingly important as the field seeks to improve the performance of existing models by scaling them up, while simultaneously trying to reduce power consumption and computational footprint. Unfortunately, most existing methods for inducing performant sparse models still entail the instantiation of dense parameters, or dense gradients in the backward-pass, during training. For very large models this requirement can be prohibitive. In this work we propose Top-KAST, a method that preserves constant sparsity throughout training (in both the forward and backward-passes). We demonstrate the efficacy of our approach by showing that it performs comparably to or better than previous works when training models on the established ImageNet benchmark, whilst fully maintaining sparsity. In addition to our ImageNet results, we also demonstrate our approach in the domain of language modeling where the current best performing architectures tend to have tens of billions of parameters and scaling up does not yet seem to have saturated performance. Sparse versions of these architectures can be run with significantly fewer resources, making them more widely accessible and applicable. Furthermore, in addition to being effective, our approach is straightforward and can easily be implemented in a wide range of existing machine learning frameworks with only a few additional lines of code. We therefore hope that our contribution will help enable the broader community to explore the potential held by massive models, without incurring massive computational cost.
**********
Meta-Learning with Adaptive Hyperparameters
Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, Kyoung Mu Lee
Despite its popularity, several recent works question the effectiveness of MAML when test tasks are different from training tasks, thus suggesting various task-conditioned methodology to improve the initialization. Instead of searching for better task-aware initialization, we focus on a complementary factor in MAML framework, inner-loop optimization (or fast adaptation). Consequently, we propose a new weight update rule that greatly enhances the fast adaptation process. Specifically, we introduce a small meta-network that can adaptively generate per-step hyperparameters: learning rate and weight decay coefficients. The experimental results validate that the Adaptive Learning of hyperparameters for Fast Adaptation (ALFA) is the equally important ingredient that was often neglected in the recent few-shot learning approaches. Surprisingly, fast adaptation from random initialization with ALFA can already outperform MAML.
**********
Tight last-iterate convergence rates for no-regret learning in multi-player games
Noah Golowich, Sarath Pattathil, Constantinos Daskalakis
We study the question of obtaining last-iterate convergence rates for no-regret learning algorithms in multi-player games. We show that the optimistic gradient (OG) algorithm with a constant step-size, which is no-regret, achieves a last-iterate rate of $O(1/\sqrt{T})$ with respect to the gap function in smooth monotone games. This result addresses a question of Mertikopoulos & Zhou (2018), who asked whether extra-gradient approaches (such as OG) can be applied to achieve improved guarantees in the multi-agent learning setting. The proof of our upper bound uses a new technique centered around an adaptive choice of potential function at each iteration. We also show that the $O(1/\sqrt{T})$ rate is tight for all p-SCLI algorithms, which includes OG as a special case. As a byproduct of our lower bound analysis we additionally present a proof of a conjecture of Arjevani et al. (2015) which is more direct than previous approaches.
**********
Curvature Regularization to Prevent Distortion in Graph Embedding
Hongbin Pei, Bingzhe Wei, Kevin Chang, Chunxu Zhang, Bo Yang
Recent research on graph embedding has achieved success in various applications. Most graph embedding methods preserve the proximity in a graph into a manifold in an embedding space. We argue an important but neglected problem about this pr

oximity-preserving strategy: Graph topology patterns, while preserved well into an embedding manifold by preserving proximity, may distort in the ambient embedding Euclidean space, and hence to detect them becomes difficult for machine learning models. To address the problem, we propose curvature regularization, to enforce flatness for embedding manifolds, thereby preventing the distortion. We present a novel angle-based sectional curvature, termed ABS curvature, and accordingly three kinds of curvature regularization to induce flat embedding manifolds during graph embedding. We integrate curvature regularization into five popular proximity-preserving embedding methods, and empirical results in two applications show significant improvements on a wide range of open graph datasets.
**********

Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability
Nathan Inkawhich, Kevin Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, Yiran 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.
**********

Statistical and Topological Properties of Sliced Probability Divergences
Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, Umut Simsekli
The idea of slicing divergences has been proven to be successful when comparing two probability measures in various machine learning applications including generative modeling, and consists in computing the expected value of a `base divergence' between \emph{one-dimensional random projections} of the two measures. However, the topological, statistical, and computational consequences of this technique have not yet been well-established. In this paper, we aim at bridging this gap and derive various theoretical properties of sliced probability divergences. First, we show that slicing preserves the metric axioms and the weak continuity of the divergence, implying that the sliced divergence will share similar topological properties. We then precise the results in the case where the base divergence belongs to the class of integral probability metrics. On the other hand, we establish that, under mild conditions, the sample complexity of a sliced divergence does not depend on the problem dimension. We finally apply our general results to several base divergences, and illustrate our theory on both synthetic and real data experiments.
**********

Probabilistic Active Meta-Learning
Jean Kaddour, Steindor Saemundsson, Marc Deisenroth (he/him)
Data-efficient learning algorithms are essential in many practical applications where data collection is expensive, e.g., in robotics due to the wear and tear. To address this problem, meta-learning algorithms use prior experience about tasks to learn new, related tasks efficiently. Typically, a set of training tasks is assumed given or randomly chosen. However, this setting does not take into account the sequential nature that naturally arises when training a model from scratch in real-life: how do we collect a set of training tasks in a data-efficient manner? In this work, we introduce task selection based on prior experience into a meta-learning algorithm by conceptualizing the learner and the active meta-learning setting using a probabilistic latent variable model. We provide empirical evidence that our approach improves data-efficiency when compared to strong baselines on simulated robotic experiments.
**********

Knowledge Distillation in Wide Neural Networks: Risk Bound, Data Efficiency and Imperfect Teacher
Guangda Ji, Zhanxing Zhu
Knowledge distillation is a strategy of training a student network with guide of the soft output from a teacher network. It has been a successful method of model compression and knowledge transfer. However, currently knowledge distillation

lacks a convincing theoretical understanding. On the other hand, recent finding on neural tangent kernel enables us to approximate a wide neural network with a linear model of the network's random features. In this paper, we theoretically analyze the knowledge distillation of a wide neural network. First we provide a transfer risk bound for the linearized model of the network. Then we propose a metric of the task's training difficulty, called data inefficiency. Based on this metric, we show that for a perfect teacher, a high ratio of teacher's soft labels can be beneficial. Finally, for the case of imperfect teacher, we find that hard labels can correct teacher's wrong prediction, which explains the practice of mixing hard and soft labels.

**********

Adversarial Attacks on Deep Graph Matching

Zijie Zhang, Zeru Zhang, Yang Zhou, Yelong Shen, Ruoming Jin, Dejing Dou

Despite achieving remarkable performance, deep graph learning models, such as node classification and network embedding, suffer from harassment caused by small adversarial perturbations. However, the vulnerability analysis of graph matching under adversarial attacks has not been fully investigated yet. This paper proposes an adversarial attack model with two novel attack techniques to perturb the graph structure and degrade the quality of deep graph matching: (1) a kernel density estimation approach is utilized to estimate and maximize node densities to derive imperceptible perturbations, by pushing attacked nodes to dense regions in two graphs, such that they are indistinguishable from many neighbors; and (2) a meta learning-based projected gradient descent method is developed to well choose attack starting points and to improve the search performance for producing effective perturbations. We evaluate the effectiveness of the attack model on real datasets and validate that the attacks can be transferable to other graph learning models.

**********

The Generalization-Stability Tradeoff In Neural Network Pruning

Brian Bartoldson, Ari Morcos, Adrian Barbu, Gordon Erlebacher

Pruning neural network parameters is often viewed as a means to compress models, but pruning has also been motivated by the desire to prevent overfitting. This motivation is particularly relevant given the perhaps surprising observation that a wide variety of pruning approaches increase test accuracy despite sometimes massive reductions in parameter counts. To better understand this phenomenon, we analyze the behavior of pruning over the course of training, finding that pruning's benefit to generalization increases with pruning's instability (defined as the drop in test accuracy immediately following pruning). We demonstrate that this "generalization-stability tradeoff'' is present across a wide variety of pruning settings and propose a mechanism for its cause: pruning regularizes similarly to noise injection. Supporting this, we find less pruning stability leads to more model flatness and the benefits of pruning do not depend on permanent parameter removal. These results explain the compatibility of pruning-based generalization improvements and the high generalization recently observed in overparameterized networks.

**********

Gradient-EM Bayesian Meta-Learning

Yayi Zou, Xiaoqi Lu

Bayesian meta-learning enables robust and fast adaptation to new tasks with uncertainty assessment. The key idea behind Bayesian meta-learning is empirical Bayes inference of hierarchical model. In this work, we extend this framework to include a variety of existing methods, before proposing our variant based on gradient-EM algorithm. Our method improves computational efficiency by avoiding back-propagation computation in the meta-update step, which is exhausting for deep neural networks. Furthermore, it provides flexibility to the inner-update optimization procedure by decoupling it from meta-update. Experiments on sinusoidal regression, few-shot image classification, and policy-based reinforcement learning show that our method not only achieves better accuracy with less computation cost, but is also more robust to uncertainty.

**********

Logarithmic Regret Bound in Partially Observable Linear Dynamical Systems
Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, Anima Anandkumar
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Linearly Converging Error Compensated SGD
Eduard Gorbunov, Dmitry Kovalev, Dmitry Makarenko, Peter Richtarik
In this paper, we propose a unified analysis of variants of distributed SGD with arbitrary compressions and delayed updates. Our framework is general enough to cover different variants of quantized SGD, Error-Compensated SGD (EC-SGD), and SGD with delayed updates (D-SGD). Via single theorem, we derive the complexity results for all the methods that fit our framework. For the existing methods, this theorem gives the best-known complexity results. Moreover, using our general scheme, we develop new variants of SGD that combine variance reduction or arbitrary sampling with error feedback and quantization and derive the convergence rates for these methods beating the state-of-the-art results. In order to illustrate the strength of our framework, we develop 16 new methods that fit this. In particular, we propose the first method called EC-SGD-DIANA that is based on error-feedback for biased compression operator and quantization of gradient differences and prove the convergence guarantees showing that EC-SGD-DIANA converges to the exact optimum asymptotically in expectation with constant learning rate for both convex and strongly convex objectives when workers compute full gradients of their loss functions. Moreover, for the case when the loss function of the worker has the form of finite sum, we modified the method and got a new one called EC-LSVRG-DIANA which is the first distributed stochastic method with error feedback and variance reduction that converges to the exact optimum asymptotically in expectation with constant learning rate.
**********

Canonical 3D Deformer Maps: Unifying parametric and non-parametric methods for dense weakly-supervised category reconstruction
David Novotny, Roman Shapovalov, Andrea Vedaldi
We propose the Canonical 3D Deformer Map, a new representation of the 3D shape of common object categories that can be learned from a collection of 2D images of independent objects. Our method builds in a novel way on concepts from parametric deformation models, non-parametric 3D reconstruction, and canonical embeddings, combining their individual advantages. In particular, it learns to associate each image pixel with a deformation model of the corresponding 3D object point which is canonical, i.e. intrinsic to the identity of the point and shared across objects of the category. The result is a method that, given only sparse 2D supervision at training time, can, at test time, reconstruct the 3D shape and texture of objects from single views, while establishing meaningful dense correspondences between object instances. It also achieves state-of-the-art results in dense 3D reconstruction on public in-the-wild datasets of faces, cars, and birds.
**********

A Self-Tuning Actor-Critic Algorithm
Tom Zahavy, Zhongwen Xu, Vivek Veeriah, Matteo Hessel, Junhyuk Oh, Hado P. van Hasselt, David Silver, Satinder Singh
Reinforcement learning algorithms are highly sensitive to the choice of hyperparameters, typically requiring significant manual effort to identify hyperparameters that perform well on a new domain. In this paper, we take a step towards addressing this issue by using metagradients to automatically adapt hyperparameters online by meta-gradient descent (Xu et al., 2018). We apply our algorithm, Self-Tuning Actor-Critic (STAC), to self-tune all the differentiable hyperparameters of an actor-critic loss function, to discover auxiliary tasks, and to improve off-policy learning using a novel leaky V-trace operator. STAC is simple to use, sample efficient and does not require a significant increase in compute. Ablative studies show that the overall performance of STAC improved as we adapt more hyperparameters. When applied to the Arcade Learning Environment (Bellemare et al.

2012), STAC improved the median human normalized score in 200M steps from 243% t
o 364%. When applied to the DM Control suite (Tassa et al., 2018), STAC improved
 the mean score in 30M steps from 217 to 389 when learning with features, from 1
08 to 202 when learning from pixels, and from 195 to 295 in the Real-World Reinf
orcement Learning Challenge (Dulac-Arnold et al., 2020).
**********

The Cone of Silence: Speech Separation by Localization
Teerapat Jenrungrot, Vivek Jayaram, Steve Seitz, Ira Kemelmacher-Shlizerman
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

High-Dimensional Bayesian Optimization via Nested Riemannian Manifolds
Noémie Jaquier, Leonel Rozo
Despite the recent success of Bayesian optimization (BO) in a variety of applica
tions where sample efficiency is imperative, its performance may be seriously co
mpromised in settings characterized by high-dimensional parameter spaces. A solu
tion to preserve the sample efficiency of BO in such problems is to introduce do
main knowledge into its formulation. In this paper, we propose to exploit the ge
ometry of non-Euclidean search spaces, which often arise in a variety of domains
, to learn structure-preserving mappings and optimize the acquisition function o
f BO in low-dimensional latent spaces. Our approach, built on Riemannian manifol
ds theory, features geometry-aware Gaussian processes that jointly learn a neste
d-manifolds embedding and a representation of the objective function in the late
nt space. We test our approach in several benchmark artificial landscapes and re
port that it not only outperforms other high-dimensional BO approaches in severa
l settings, but consistently optimizes the objective functions, as opposed to ge
ometry-unaware BO methods.
**********

Train-by-Reconnect: Decoupling Locations of Weights from Their Values
Yushi Qiu, Reiji Suda
What makes untrained deep neural networks (DNNs) different from the trained perf
ormant ones? By zooming into the weights in well-trained DNNs, we found that it
is the location of weights that holds most of the information encoded by the tra
ining. Motivated by this observation, we hypothesized that weights in DNNs train
ed using stochastic gradient-based methods can be separated into two dimensions:
 the location of weights, and their exact values. To assess our hypothesis, we p
ropose a novel method called lookahead permutation (LaPerm) to train DNNs by rec
onnecting the weights. We empirically demonstrate LaPerm's versatility while pro
ducing extensive evidence to support our hypothesis: when the initial weights ar
e random and dense, our method demonstrates speed and performance similar to or
better than that of regular optimizers, e.g., Adam. When the initial weights are
 random and sparse (many zeros), our method changes the way neurons connect, ach
ieving accuracy comparable to that of a well-trained dense network. When the ini
tial weights share a single value, our method finds a weight agnostic neural net
work with far-better-than-chance accuracy.
**********

Learning discrete distributions: user vs item-level privacy
Yuhan Liu, Ananda Theertha Suresh, Felix Xinnan X. Yu, Sanjiv Kumar, Michael Ril
ey
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Matrix Completion with Quantified Uncertainty through Low Rank Gaussian Copula
Yuxuan Zhao, Madeleine Udell
Modern large scale datasets are often plagued with missing entries. For tabular
data with missing values, a flurry of imputation algorithms solve for a complete

matrix which minimizes some penalized reconstruction error. However, almost none of them can estimate the uncertainty of its imputations. This paper pro- poses a probabilistic and scalable framework for missing value imputation with quantified uncertainty. Our model, the Low Rank Gaussian Copula, augments a standard probabilistic model, Probabilistic Principal Component Analysis, with marginal transformations for each column that allow the model to better match the distribution of the data. It naturally handles Boolean, ordinal, and real-valued observations and quantifies the uncertainty in each imputation. The time required to fit the model scales linearly with the number of rows and the number of columns in the dataset. Empirical results show the method yields state-of-the-art imputation accuracy across a wide range of data types, including those with high rank. Our uncertainty measure predicts imputation error well: entries with lower uncertainty do have lower imputation error (on average). Moreover, for real-valued data, the resulting confidence intervals are well-calibrated.
**********

Sparse and Continuous Attention Mechanisms
André Martins, António Farinhas, Marcos Treviso, Vlad Niculae, Pedro Aguiar, Mario Figueiredo
Exponential families are widely used in machine learning; they include many distributions in continuous and discrete domains (e.g., Gaussian, Dirichlet, Poisson, and categorical distributions via the softmax transformation). Distributions in each of these families have fixed support. In contrast, for finite domains, there has been recent work on sparse alternatives to softmax (e.g., sparsemax and alpha-entmax), which have varying support, being able to assign zero probability to irrelevant categories. These discrete sparse mappings have been used for improving interpretability of neural attention mechanisms. This paper expands that work in two directions: first, we extend alpha-entmax to continuous domains, revealing a link with Tsallis statistics and deformed exponential families. Second, we introduce continuous-domain attention mechanisms, deriving efficient gradient backpropagation algorithms for alpha in {1,2}. Experiments on attention-based text classification, machine translation, and visual question answering illustrate the use of continuous attention in 1D and 2D, showing that it allows attending to time intervals and compact regions.
**********

Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection
Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, Jian Yang
One-stage detector basically formulates object detection as dense classification and localization (i.e., bounding box regression). The classification is usually optimized by Focal Loss and the box location is commonly learned under Dirac delta distribution. A recent trend for one-stage detectors is to introduce an \emph{individual} prediction branch to estimate the quality of localization, where the predicted quality facilitates the classification to improve detection performance. This paper delves into the \emph{representations} of the above three fundamental elements: quality estimation, classification and localization. Two problems are discovered in existing practices, including (1) the inconsistent usage of the quality estimation and classification between training and inference, and (2) the inflexible Dirac delta distribution for localization. To address the problems, we design new representations for these elements. Specifically, we merge the quality estimation into the class prediction vector to form a joint representation, and use a vector to represent arbitrary distribution of box locations. The improved representations eliminate the inconsistency risk and accurately depict the flexible distribution in real data, but contain \emph{continuous} labels, which is beyond the scope of Focal Loss. We then propose Generalized Focal Loss (GFL) that generalizes Focal Loss from its discrete form to the \emph{continuous} version for successful optimization. On COCO {\tt test-dev}, GFL achieves 45.0\% AP using ResNet-101 backbone, surpassing state-of-the-art SAPD (43.5\%) and ATSS (43.6\%) with higher or comparable inference speed.
**********

Learning by Minimizing the Sum of Ranked Range
Shu Hu, Yiming Ying, xin wang, Siwei Lyu
In forming learning objectives, one oftentimes needs to aggregate a set of individual values to a single output. Such cases occur in the aggregate loss, which combines individual losses of a learning model over each training sample, and in the individual loss for multi-label learning, which combines prediction scores over all class labels. In this work, we introduce the sum of ranked range (SoRR) as a general approach to form learning objectives. A ranked range is a consecutive sequence of sorted values of a set of real numbers. The minimization of SoRR is solved with the difference of convex algorithm (DCA). We explore two applications in machine learning of the minimization of the SoRR framework, namely the AoRR aggregate loss for binary classification and the TKML individual loss for multi-label/multi-class classification. Our empirical results highlight the effectiveness of the proposed optimization framework and demonstrate the applicability of proposed losses using synthetic and real datasets.
**********
Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations
Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, Cho-Ju i Hsieh
A deep reinforcement learning (DRL) agent observes its states through observations, which may contain natural measurement errors or adversarial noises. Since the observations deviate from the true states, they can mislead the agent into making suboptimal actions. Several works have shown this vulnerability via adversarial attacks, but how to improve the robustness of DRL under this setting has not been well studied. We show that naively applying existing techniques on improving robustness for classification tasks, like adversarial training, are ineffective for many RL tasks. We propose the state-adversarial Markov decision process (SA-MDP) to study the fundamental properties of this problem, and develop a theoretically principled policy regularization which can be applied to a large family of DRL algorithms, including deep deterministic policy gradient (DDPG), proximal policy optimization (PPO) and deep Q networks (DQN), for both discrete and continuous action control problems. We significantly improve the robustness of DDPG, PPO and DQN agents under a suite of strong white box adversarial attacks, including two new attacks of our own. Additionally, we find that a robust policy noticeably improves DRL performance in a number of environments.
**********
Understanding Anomaly Detection with Deep Invertible Networks through Hierarchies of Distributions and Features
Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, Dan Zhang
Deep generative networks trained via maximum likelihood on a natural image dataset like CIFAR10 often assign high likelihoods to images from datasets with different objects (e.g., SVHN). We refine previous investigations of this failure at anomaly detection for invertible generative networks and provide a clear explanation of it as a combination of model bias and domain prior: Convolutional networks learn similar low-level feature distributions when trained on any natural image dataset and these low-level features dominate the likelihood. Hence, when the discriminative features between inliers and outliers are on a high-level, e.g., object shapes, anomaly detection becomes particularly challenging. To remove the negative impact of model bias and domain prior on detecting high-level differences, we propose two methods, first, using the log likelihood ratios of two identical models, one trained on the in-distribution data (e.g., CIFAR10) and the other one on a more general distribution of images (e.g., 80 Million Tiny Images). We also derive a novel outlier loss for the in-distribution network on samples from the more general distribution to further improve the performance. Secondly, using a multi-scale model like Glow, we show that low-level features are mainly captured at early scales. Therefore, using only the likelihood contribution of the final scale performs remarkably well for detecting high-level feature differences of the out-of-distribution and the in-distribution. This method is especially useful if one does not have access to a suitable general distribution. Over

all, our methods achieve strong anomaly detection performance in the unsupervise d setting, and only slightly underperform state-of-the-art classifier-based meth ods in the supervised setting. Code can be found at https://github.com/boschrese arch/hierarchicalanomalydetection.
**********

Fair Hierarchical Clustering
Sara Ahmadian, Alessandro Epasto, Marina Knittel, Ravi Kumar, Mohammad Mahdian, Benjamin Moseley, Philip Pham, Sergei Vassilvitskii, Yuyan Wang
As machine learning has become more prevalent, researchers have begun to recogni ze the necessity of ensuring machine learning systems are fair. Recently, there has been an interest in defining a notion of fairness that mitigates over-repres entation in traditional clustering.
**********

Self-training Avoids Using Spurious Features Under Domain Shift
Yining Chen, Colin Wei, Ananya Kumar, Tengyu Ma
In unsupervised domain adaptation, existing theory focuses on situations where t he source and target domains are close. In practice, conditional entropy minimiz ation and pseudo-labeling work even when the domain shifts are much larger than those analyzed by existing theory. We identify and analyze one particular settin g where the domain shift can be large, but these algorithms provably work: certa in spurious features correlate with the label in the source domain but are indep endent of the label in the target. Our analysis considers linear classification where the spurious features are Gaussian and the non-spurious features are a mix ture of log-concave distributions. For this setting, we prove that entropy minim ization on unlabeled target data will avoid using the spurious feature if initia lized with a decently accurate source classifier, even though the objective is n on-convex and contains multiple bad local minima using the spurious features. We verify our theory for spurious domain shift tasks on semi-synthetic Celeb-A and MNIST datasets. Our results suggest that practitioners collect and self-train o n large, diverse datasets to reduce biases in classifiers even if labeling is im practical.
**********

Improving Online Rent-or-Buy Algorithms with Sequential Decision Making and ML P redictions
Shom Banerjee
In this work we study online rent-or-buy problems as a sequential decision makin g problem. We show how one can integrate predictions, typically coming from a ma chine learning (ML) setup, into this framework. Specifically, we consider the sk i-rental problem and the dynamic TCP acknowledgment problem. We present new onli ne algorithms and obtain explicit performance bounds in-terms of the accuracy of the prediction. Our algorithms are close to optimal with accurate predictions w hile hedging against less accurate predictions.
**********

CircleGAN: Generative Adversarial Learning across Spherical Circles
Woohyeon Shim, Minsu Cho
We present a novel discriminator for GANs that improves realness and diversity o f generated samples by learning a structured hypersphere embedding space using s pherical circles.
The proposed discriminator learns to populate realistic samples around the longe st spherical circle, i.e., a great circle, while pushing unrealistic samples tow ard the poles perpendicular to the great circle. Since longer circles occupy lar ger area on the hypersphere, they encourage more diversity in representation lea rning, and vice versa. Discriminating samples based on their corresponding spher ical circles can thus naturally induce diversity to generated samples.
We also extend the proposed method for conditional settings with class labels by creating a hypersphere for each category and performing class-wise discriminati on and update. In experiments, we validate the effectiveness for both unconditio nal and conditional generation on standard benchmarks, achieving the state of th e art.
**********

WOR and $p$'s: Sketches for $\ell_p$-Sampling Without Replacement
Edith Cohen, Rasmus Pagh, 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.
**********
Hypersolvers: Toward Fast Continuous-Depth Models
Michael Poli, Stefano Massaroli, Atsushi Yamashita, Hajime Asama, Jinkyoo Park
The infinite-depth paradigm pioneered by Neural ODEs has launched a renaissance in the search for novel dynamical system-inspired deep learning primitives; however, their utilization in problems of non-trivial size has often proved impossible due to poor computational scalability. This work paves the way for scalable Neural ODEs with time-to-prediction comparable to traditional discrete networks. We introduce hypersolvers, neural networks designed to solve ODEs with low overhead and theoretical guarantees on accuracy. The synergistic combination of hypersolvers and Neural ODEs allows for cheap inference and unlocks a new frontier for practical application of continuous-depth models. Experimental evaluations on standard benchmarks, such as sampling for continuous normalizing flows, reveal consistent pareto efficiency over classical numerical methods.
**********
Log-Likelihood Ratio Minimizing Flows: Towards Robust and Quantifiable Neural Distribution Alignment
Ben Usman, Avneesh Sud, Nick Dufour, Kate Saenko
Distribution alignment has many applications in deep learning, including domain adaptation and unsupervised image-to-image translation. Most prior work on unsupervised distribution alignment relies either on minimizing simple non-parametric statistical distances such as maximum mean discrepancy or on adversarial alignment. However, the former fails to capture the structure of complex real-world distributions, while the latter is difficult to train and does not provide any universal convergence guarantees or automatic quantitative validation procedures. In this paper, we propose a new distribution alignment method based on a log-likelihood ratio statistic and normalizing flows. We show that, under certain assumptions, this combination yields a deep neural likelihood-based minimization objective that attains a known lower bound upon convergence. We experimentally verify that minimizing the resulting objective results in domain alignment that preserves the local structure of input domains.
**********
Escaping the Gravitational Pull of Softmax
Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvari, Dale Schuurmans
The softmax is the standard transformation used in machine learning to map real-valued vectors to categorical distributions. Unfortunately, this transform poses serious drawbacks for gradient descent (ascent) optimization. We reveal this difficulty by establishing two negative results: (1) optimizing any expectation with respect to the softmax must exhibit sensitivity to parameter initialization (softmax gravity well''), and (2) optimizing log-probabilities under the softmax must exhibit slow convergence (softmax damping''). Both findings are based on an analysis of convergence rates using the Non-uniform \L{}ojasiewicz (N\L{}) inequalities. To circumvent these shortcomings we investigate an alternative transformation, the \emph{escort} mapping, that demonstrates better optimization properties. The disadvantages of the softmax and the effectiveness of the escort transformation are further explained using the concept of N\L{} coefficient. In addition to proving bounds on convergence rates to firmly establish these results, we also provide experimental evidence for the superiority of the escort transformation.
**********
Regret in Online Recommendation Systems
Kaito Ariu, Narae Ryu, Se-Young Yun, Alexandre Proutiere
Requests for name changes in the electronic proceedings will be accepted with 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 Convergence and Generalization of Dropout Training
Poorya Mianjy, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Second Order Optimality in Decentralized Non-Convex Optimization via Perturbed G
radient Tracking
Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari
In this paper we study the problem of escaping from saddle points and achieving
 second-order optimality in a decentralized setting where a group of agents colla
borate to minimize their aggregate objective function. We provide a non-asymptot
ic (finite-time) analysis and show that by following the idea of perturbed gradi
ent descent, it is possible to converge to a second-order stationary point in a
number of iterations which depends linearly on dimension and polynomially on the
 accuracy of second-order stationary point. Doing this in a communication-effici
ent manner requires overcoming several challenges, from identifying (first order
) stationary points in a distributed manner, to adapting the perturbed gradient
framework without prohibitive communication complexity. Our proposed Perturbed D
ecentralized Gradient Tracking (PDGT) method consists of two major stages: (i) a
 gradient-based step to find a first-order stationary point and (ii) a perturbed
 gradient descent step to escape from a first-order stationary point, if it is a
 saddle point with sufficient curvature. As a side benefit of our result, in the
 case that all saddle points are non-degenerate (strict), the proposed PDGT meth
od finds a local minimum of the considered decentralized optimization problem in
 a finite number of iterations.
**********

Implicit Regularization in Deep Learning May Not Be Explainable by Norms
Noam Razin, Nadav Cohen
Mathematically characterizing the implicit regularization induced by gradient-ba
sed optimization is a longstanding pursuit in the theory of deep learning. A wid
espread hope is that a characterization based on minimization of norms may apply
, and a standard test-bed for studying this prospect is matrix factorization (ma
trix completion via linear neural networks). It is an open question whether norm
s can explain the implicit regularization in matrix factorization. The current p
aper resolves this open question in the negative, by proving that there exist na
tural matrix factorization problems on which the implicit regularization drives
all norms (and quasi-norms) towards infinity. Our results suggest that, rather t
han perceiving the implicit regularization via norms, a potentially more useful
interpretation is minimization of rank. We demonstrate empirically that this int
erpretation extends to a certain class of non-linear neural networks, and hypoth
esize that it may be key to explaining generalization in deep learning.
**********

POMO: Policy Optimization with Multiple Optima for Reinforcement Learning
Yeong-Dae Kwon, Jinho Choo, Byoungjip Kim, Iljoo Yoon, Youngjune Gwon, Seungjai
Min
In neural combinatorial optimization (CO), reinforcement learning (RL) can turn
a deep neural net into a fast, powerful heuristic solver of NP-hard problems. Th
is approach has a great potential in practical applications because it allows ne
ar-optimal solutions to be found without expert guides armed with substantial do
main knowledge. We introduce Policy Optimization with Multiple Optima (POMO), an
 end-to-end approach for building such a heuristic solver. POMO is applicable to
 a wide range of CO problems. It is designed to exploit the symmetries in the re
presentation of a CO solution. POMO uses a modified REINFORCE algorithm that for
ces diverse rollouts towards all optimal solutions. Empirically, the low-varianc
e baseline of POMO makes RL training fast and stable, and it is more resistant t

o local minima compared to previous approaches. We also introduce a new augmenta tion-based inference method, which accompanies POMO nicely. We demonstrate the e ffectiveness of POMO by solving three popular NP-hard problems, namely, travelin g salesman (TSP), capacitated vehicle routing (CVRP), and 0-1 knapsack (KP). For all three, our solver based on POMO shows a significant improvement in performa nce over all recent learned heuristics. In particular, we achieve the optimality gap of 0.14% with TSP100 while reducing inference time by more than an order of magnitude.
**********
Uncertainty-aware Self-training for Few-shot Text Classification
Subhabrata Mukherjee, Ahmed Awadallah
Recent success of pre-trained language models crucially hinges on fine-tuning th em on large amounts of labeled data for the downstream task, that are typically expensive to acquire or difficult to access for many applications. We study self -training as one of the earliest semi-supervised learning approaches to reduce t he annotation bottleneck by making use of large-scale unlabeled data for the tar get task. Standard self-training mechanism randomly samples instances from the u nlabeled pool to generate pseudo-labels and augment labeled data. We propose an approach to improve self-training by incorporating uncertainty estimates of the underlying neural network leveraging recent advances in Bayesian deep learning. Specifically, we propose (i) acquisition functions to select instances from the unlabeled pool leveraging Monte Carlo (MC) Dropout, and (ii) learning mechanism leveraging model confidence for self-training. As an application, we focus on te xt classification with five benchmark datasets. We show our methods leveraging o nly 20-30 labeled samples per class for each task for training and for validatio n perform within 3% of fully supervised pre-trained language models fine-tuned o n thousands of labels with an aggregate accuracy of 91% and improvement of up to 12% over baselines.
**********
Learning to Learn with Feedback and Local Plasticity
Jack Lindsey, Ashok Litwin-Kumar
Interest in biologically inspired alternatives to backpropagation is driven by t he desire to both advance connections between deep learning and neuroscience and address backpropagation's shortcomings on tasks such as online, continual learn ing. However, local synaptic learning rules like those employed by the brain hav e so far failed to match the performance of backpropagation in deep networks. I n this study, we employ meta-learning to discover networks that learn using feed back connections and local, biologically inspired learning rules. Importantly, the feedback connections are not tied to the feedforward weights, avoiding biolo gically implausible weight transport. Our experiments show that meta-trained ne tworks effectively use feedback connections to perform online credit assignment in multi-layer architectures. Surprisingly, this approach matches or exceeds a state-of-the-art gradient-based online meta-learning algorithm on regression and classification tasks, excelling in particular at continual learning. Analysis of the weight updates employed by these models reveals that they differ qualitat ively from gradient descent in a way that reduces interference between updates. Our results suggest the existence of a class of biologically plausible learning mechanisms that not only match gradient descent-based learning, but also overco me its limitations.
**********
Every View Counts: Cross-View Consistency in 3D Object Detection with Hybrid-Cyl indrical-Spherical Voxelization
Qi Chen, Lin Sun, Ernest Cheung, Alan L. Yuille
Recent voxel-based 3D object detectors for autonomous vehicles learn point cloud representations either from bird eye view (BEV) or range view (RV, a.k.a. the p erspective view). However, each view has its own strengths and weaknesses. In th is paper, we present a novel framework to unify and leverage the benefits from b oth BEV and RV. The widely-used cuboid-shaped voxels in Cartesian coordinate sys tem only benefit learning BEV feature map. Therefore, to enable learning both BE V and RV feature maps, we introduce Hybrid-Cylindrical-Spherical voxelization. O

ur findings show that simply adding detection on another view as auxiliary super vision will lead to poor performance. We proposed a pair of cross-view transform ers to transform the feature maps into the other view and introduce cross-view c onsistency loss on them. Comprehensive experiments on the challenging NuScenes D ataset validate the effectiveness of our proposed method by virtue of joint opti mization and complementary information on both views. Remarkably, our approach a chieved mAP of 55.8%, outperforming all published approaches by at least 3% in o verall performance and up to 16.5% in safety-crucial categories like cyclist.
**********
Sharper Generalization Bounds for Pairwise Learning
Yunwen Lei, Antoine Ledent, Marius Kloft
Requests for name changes in the electronic proceedings will be accepted with 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 Measure-Theoretic Approach to Kernel Conditional Mean Embeddings
Junhyung Park, Krikamol Muandet
We present a new operator-free, measure-theoretic approach to the conditional me an embedding as a random variable taking values in a reproducing kernel Hilbert space. While the kernel mean embedding of marginal distributions has been define d rigorously, the existing operator-based approach of the conditional version la cks a rigorous treatment, and depends on strong assumptions that hinder its anal ysis. Our approach does not impose any of the assumptions that the operator-base d counterpart requires. We derive a natural regression interpretation to obtain empirical estimates, and provide a thorough analysis of its properties, includin g universal consistency with improved convergence rates. As natural by-products, we obtain the conditional analogues of the Maximum Mean Discrepancy and Hilbert -Schmidt Independence Criterion, and demonstrate their behaviour via simulations .
**********
Quantifying the Empirical Wasserstein Distance to a Set of Measures: Beating the Curse of Dimensionality
Nian Si, Jose Blanchet, Soumyadip Ghosh, Mark Squillante
We consider the problem of estimating the Wasserstein distance between the empir ical measure and a set of probability measures whose expectations over a class o f functions (hypothesis class) are constrained. If this class is sufficiently ri ch to characterize a particular distribution (e.g., all Lipschitz functions), th en our formulation recovers the Wasserstein distance to such a distribution. We establish a strong duality result that generalizes the celebrated Kantorovich-Ru binstein duality. We also show that our formulation can be used to beat the curs e of dimensionality, which is well known to affect the rates of statistical conv ergence of the empirical Wasserstein distance. In particular, examples of infini te-dimensional hypothesis classes are presented, informed by a complex correlati on structure, for which it is shown that the empirical Wasserstein distance to s uch classes converges to zero at the standard parametric rate. Our formulation p rovides insights that help clarify why, despite the curse of dimensionality, the Wasserstein distance enjoys favorable empirical performance across a wide range of statistical applications.
**********
Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning
Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Riche mond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohamm ad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, Michal Valko
We introduce Bootstrap Your Own Latent (BYOL), a new approach to self-supervised image representation learning. BYOL relies on two neural networks, referred to as online and target networks, that interact and learn from each other. From an augmented view of an image, we train the online network to predict the target ne twork representation of the same image under a different augmented view. At the same time, we update the target network with a slow-moving average of the online

network. While state-of-the art methods intrinsically rely on negative pairs, BYOL achieves a new state of the art without them. BYOL reaches 74.3% top-1 classification accuracy on ImageNet using the standard linear evaluation protocol with a standard ResNet-50 architecture and 79.6% with a larger ResNet. We also show that BYOL performs on par or better than the current state of the art on both transfer and semi-supervised benchmarks.
**********
Towards Theoretically Understanding Why Sgd Generalizes Better Than Adam in Deep Learning
Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, Weinan E
It is not clear yet why ADAM-alike adaptive gradient algorithms suffer from worse generalization performance than SGD despite their faster training speed. This work aims to provide understandings on this generalization gap by analyzing their local convergence behaviors. Specifically, we observe the heavy tails of gradient noise in these algorithms. This motivates us to analyze these algorithms through their Levy-driven stochastic differential equations (SDEs) because of the similar convergence behaviors of an algorithm and its SDE. Then we establish the escaping time of these SDEs from a local basin. The result shows that (1) the escaping time of both SGD and ADAM~depends on the Radon measure of the basin positively and the heaviness of gradient noise negatively; (2) for the same basin, SGD enjoys smaller escaping time than ADAM, mainly because (a) the geometry adaptation in ADAM~via adaptively scaling each gradient coordinate well diminishes the anisotropic structure in gradient noise and results in larger Radon measure of a basin; (b) the exponential gradient average in ADAM~smooths its gradient and leads to lighter gradient noise tails than SGD. So SGD is more locally unstable than ADAM~at sharp minima defined as the minima whose local basins have small Radon measure, and can better escape from them to flatter ones with larger Radon measure. As flat minima here which often refer to the minima at flat or asymmetric basins/valleys often generalize better than sharp ones~\cite{keskar2016large,he2019asymmetric}, our result explains the better generalization performance of SGD over ADAM. Finally, experimental results confirm our heavy-tailed gradient noise assumption and theoretical affirmation.
**********
RSKDD-Net: Random Sample-based Keypoint Detector and Descriptor
Fan Lu, Guang Chen, Yinlong Liu, Zhongnan Qu, Alois Knoll
Keypoint detector and descriptor are two main components of point cloud registration. Previous learning-based keypoint detectors rely on saliency estimation for each point or farthest point sample (FPS) for candidate points selection, which are inefficient and not applicable in large scale scenes. This paper proposes Random Sample-based Keypoint Detector and Descriptor Network (RSKDD-Net) for large scale point cloud registration. The key idea is using random sampling to efficiently select candidate points and using a learning-based method to jointly generate keypoints and corresponding descriptors. To tackle the information loss of random sampling, we exploit a novel random dilation cluster strategy to enlarge the receptive field of each sampled point and an attention mechanism to aggregate the positions and features of neighbor points. Furthermore, we propose a matching loss to train the descriptor in a weakly supervised manner. Extensive experiments on two large scale outdoor LiDAR datasets show that the proposed RSKDD-Net achieves state-of-the-art performance with more than 15 times faster than existing methods. Our code is available at https://github.com/ispc-lab/RSKDD-Net.
**********
Efficient Clustering for Stretched Mixtures: Landscape and Optimality
Kaizheng Wang, Yuling Yan, Mateo Diaz
This paper considers a canonical clustering problem where one receives unlabeled samples drawn from a balanced mixture of two elliptical distributions and aims for a classifier to estimate the labels. Many popular methods including PCA and k-means require individual components of the mixture to be somewhat spherical, and perform poorly when they are stretched. To overcome this issue, we propose a non-convex program seeking for an affine transform to turn the data into a one-dimensional point cloud concentrating around -1 and 1, after which clustering bec

omes easy. Our theoretical contributions are two-fold: (1) we show that the non-convex loss function exhibits desirable geometric properties when the sample size exceeds some constant multiple of the dimension, and (2) we leverage this to prove that an efficient first-order algorithm achieves near-optimal statistical precision without good initialization. We also propose a general methodology for clustering with flexible choices of feature transforms and loss objectives.
**********

A Group-Theoretic Framework for Data Augmentation
Shuxiao Chen, Edgar Dobriban, Jane Lee
Data augmentation has become an important part of modern deep learning pipelines and is typically needed to achieve state of the art performance for many learning tasks. It utilizes invariant transformations of the data, such as rotation, scale, and color shift, and the transformed images are added to the training set. However, these transformations are often chosen heuristically and a clear theoretical framework to explain the performance benefits of data augmentation is not available. In this paper, we develop such a framework to explain data augmentation as averaging over the orbits of the group that keeps the data distribution approximately invariant, and show that it leads to variance reduction. We study finite-sample and asymptotic empirical risk minimization and work out as examples the variance reduction in certain two-layer neural networks. We further propose a strategy to exploit the benefits of data augmentation for general learning tasks.
**********

The Statistical Cost of Robust Kernel Hyperparameter Turning
Raphael Meyer, Christopher Musco
This paper studies the statistical complexity of kernel hyperparameter tuning in the setting of active regression under adversarial noise. We consider the problem of finding the best interpolant from a class of kernels with unknown hyperparameters, assuming only that the noise is square-integrable. We provide finite-sample guarantees for the problem, characterizing how increasing the complexity of the kernel class increases the complexity of learning kernel hyperparameters. For common kernel classes (e.g. squared-exponential kernels with unknown lengthscale), our results show that hyperparameter optimization increases sample complexity by just a logarithmic factor, in comparison to the setting where optimal parameters are known in advance. Our result is based on a subsampling guarantee for linear regression under multiple design matrices which may be of independent interest.
**********

How does Weight Correlation Affect Generalisation Ability of Deep Neural Networks?
Gaojie Jin, Xinping Yi, Liang Zhang, Lijun Zhang, Sven Schewe, Xiaowei Huang
This paper studies the novel concept of weight correlation in deep neural networks and discusses its impact on the networks' generalisation ability. For fully-connected layers, the weight correlation is defined as the average cosine similarity between weight vectors of neurons, and for convolutional layers, the weight correlation is defined as the cosine similarity between filter matrices. Theoretically, we show that, weight correlation can, and should, be incorporated into the PAC Bayesian framework for the generalisation of neural networks, and the resulting generalisation bound is monotonic with respect to the weight correlation. We formulate a new complexity measure,  which lifts the PAC Bayes measure with weight correlation, and experimentally confirm that it is able to rank the generalisation errors of a set of networks more precisely than existing measures. More importantly, we develop a new regulariser for training, and provide extensive experiments that show that the generalisation error can be greatly reduced with our novel approach.
**********

ContraGAN: Contrastive Learning for Conditional Image Generation
Minguk Kang, Jaesik Park
Conditional image generation is the task of generating diverse images using class label information. Although many conditional Generative Adversarial Networks (

GAN) have shown realistic results, such methods consider pairwise relations betw
een the embedding of an image and the embedding of the corresponding label (data
-to-class relations) as the conditioning losses. In this paper, we propose Contr
aGAN that considers relations between multiple image embeddings in the same batc
h (data-to-data relations) as well as the data-to-class relations by using a con
ditional contrastive loss. The discriminator of ContraGAN discriminates the auth
enticity of given samples and minimizes a contrastive objective to learn the rel
ations between training images. Simultaneously, the generator tries to generate
realistic images that deceive the authenticity and have a low contrastive loss.
The experimental results show that ContraGAN outperforms state-of-the-art-models
by 7.3% and 7.7% on Tiny ImageNet and ImageNet datasets, respectively. Besides,
we experimentally demonstrate that ContraGAN helps to relieve the overfitting o
f the discriminator. For a fair comparison, we re-implement twelve state-of-the-
art GANs using the PyTorch library. The software package is available at https:/
/github.com/POSTECH-CVLab/PyTorch-StudioGAN.
**********

On the distance between two neural networks and the stability of learning
Jeremy Bernstein, Arash Vahdat, Yisong Yue, Ming-Yu Liu
This paper relates parameter distance to gradient breakdown for a broad class of
nonlinear compositional functions. The analysis leads to a new distance functio
n called deep relative trust and a descent lemma for neural networks. Since the
resulting learning rule seems to require little to no learning rate tuning, it m
ay unlock a simpler workflow for training deeper and more complex neural network
s. The Python code used in this paper is here: https://github.com/jxbz/fromage.
**********

A Topological Filter for Learning with Label Noise
Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, Chao Chen
Noisy labels can impair the performance of deep neural networks. To tackle this
problem, in this paper, we propose a new method for filtering label noise. Unlik
e most existing methods relying on the posterior probability of a noisy classifi
er, we focus on the much richer spatial behavior of data in the latent represent
ational space. By leveraging the high-order topological information of data, we
are able to collect most of the clean data and train a high-quality model. Theor
etically we prove that this topological approach is guaranteed to collect the cl
ean data with high probability. Empirical results show that our method outperfor
ms the state-of-the-arts and is robust to a broad spectrum of noise types and le
vels.
**********

Personalized Federated Learning with Moreau Envelopes
Canh T. Dinh, Nguyen Tran, Josh Nguyen
Federated learning (FL) is a decentralized and privacy-preserving machine learni
ng technique in which a group of clients collaborate with a server to learn a g
lobal model without sharing clients' data. One challenge associated with FL is s
tatistical diversity among clients, which restricts the global model from delive
ring good performance on each client's task. To address this, we propose an algo
rithm for personalized FL (pFedMe) using Moreau envelopes as clients' regulariz
ed loss functions, which help decouple personalized model optimization from the
global model learning in a bi-level problem stylized for personalized FL. Theo
retically, we show that pFedMe convergence rate is state-of-the-art: achieving
quadratic speedup for strongly convex and sublinear speedup of order 2/3 for smo
oth nonconvex objectives. Experimentally, we verify that pFedMe excels at empir
ical performance compared with the vanilla FedAvg and Per-FedAvg, a meta-learnin
g based personalized FL algorithm.
**********

Avoiding Side Effects in Complex Environments
Alex Turner, Neale Ratzlaff, Prasad Tadepalli
Reward function specification can be difficult. Rewarding the agent for making a
widget may be easy, but penalizing the multitude of possible negative side effe
cts is hard. In toy environments, Attainable Utility Preservation (AUP) avoided
side effects by penalizing shifts in the ability to achieve randomly generated

goals. We scale this approach to large, randomly generated environments based on Conway's Game of Life. By preserving optimal value for a single randomly generated reward function, AUP incurs modest overhead while leading the agent to complete the specified task and avoid many side effects. Videos and code are available at https://avoiding-side-effects.github.io/.

**********

No-regret Learning in Price Competitions under Consumer Reference Effects

Negin Golrezaei, Patrick Jaillet, Jason Cheuk Nam Liang

We study long-run market stability for repeated price competitions between two firms, where consumer demand depends on firms' posted prices and consumers' price expectations called reference prices. Consumers' reference prices vary over time according to a memory-based dynamic, which is a weighted average of all historical prices. We focus on the setting where firms are not aware of demand functions and how reference prices are formed but have access to an oracle that provides a measure of consumers' responsiveness to the current posted prices. We show that if the firms run no-regret algorithms, in particular, online mirror descent (OMD), with decreasing step sizes, the market stabilizes in the sense that firms' prices and reference prices converge to a stable Nash Equilibrium (SNE). Interestingly, we also show that there exist constant step sizes under which the market stabilizes. We further characterize the rate of convergence to the SNE for both decreasing and constant OMD step sizes.

**********

Geometric Dataset Distances via Optimal Transport

David Alvarez-Melis, Nicolo Fusi

The notion of task similarity is at the core of various machine learning paradigms, such as domain adaptation and meta-learning. Current methods to quantify it are often heuristic, make strong assumptions on the label sets across the tasks, and many are architecture-dependent, relying on task-specific optimal parameters (e.g., require training a model on each dataset). In this work we propose an alternative notion of distance between datasets that (i) is model-agnostic, (ii) does not involve training, (iii) can compare datasets even if their label sets are completely disjoint and (iv) has solid theoretical footing. This distance relies on optimal transport, which provides it with rich geometry awareness, interpretable correspondences and well-understood properties. Our results show that this novel distance provides meaningful comparison of datasets, and correlates well with transfer learning hardness across various experimental settings and datasets.

**********

Task-Agnostic Amortized Inference of Gaussian Process Hyperparameters

Sulin Liu, Xingyuan Sun, Peter J. Ramadge, Ryan P. Adams

Gaussian processes (GPs) are flexible priors for modeling functions. However, their success depends on the kernel accurately reflecting the properties of the data. One of the appeals of the GP framework is that the marginal likelihood of the kernel hyperparameters is often available in closed form, enabling optimization and sampling procedures to fit these hyperparameters to data. Unfortunately, point-wise evaluation of the marginal likelihood is expensive due to the need to solve a linear system; searching or sampling the space of hyperparameters thus often dominates the practical cost of using GPs. We introduce an approach to the identification of kernel hyperparameters in GP regression and related problems that sidesteps the need for costly marginal likelihoods. Our strategy is to "amortize" inference over hyperparameters by training a single neural network, which consumes a set of regression data and produces an estimate of the kernel function, useful across different tasks. To accommodate the varying dimension and cardinality of different regression problems, we use a hierarchical self-attention-based neural network that produces estimates of the hyperparameters which are invariant to the order of the input data points and data dimensions. We show that a single neural model trained on synthetic data is able to generalize directly to several different unseen real-world GP use cases. Our experiments demonstrate that the estimated hyperparameters are comparable in quality to those from the conventional model selection procedures, while being much faster to obtain, signifi

cantly accelerating GP regression and its related applications such as Bayesian optimization and Bayesian quadrature. The code and pre-trained model are available at https://github.com/PrincetonLIPS/AHGP.
**********

A novel variational form of the Schatten-$p$ quasi-norm
Paris Giampouras, Rene Vidal, Athanasios Rontogiannis, Benjamin Haeffele
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Energy-based Out-of-distribution Detection
Weitang Liu, Xiaoyun Wang, John Owens, Yixuan Li
Determining whether inputs are out-of-distribution (OOD) is an essential building block for safely deploying machine learning models in the open world. However, previous methods relying on the softmax confidence score suffer from overconfident posterior distributions for OOD data. We propose a unified framework for OOD detection that uses an energy score. We show that energy scores better distinguish in- and out-of-distribution samples than the traditional approach using the softmax scores. Unlike softmax confidence scores, energy scores are theoretically aligned with the probability density of the inputs and are less susceptible to the overconfidence issue. Within this framework, energy can be flexibly used as a scoring function for any pre-trained neural classifier as well as a trainable cost function to shape the energy surface explicitly for OOD detection. On a CIFAR-10 pre-trained WideResNet, using the energy score reduces the average FPR (at TPR 95%) by 18.03% compared to the softmax confidence score. With energy-based training, our method outperforms the state-of-the-art on common benchmarks.
**********

On the Loss Landscape of Adversarial Training: Identifying Challenges and How to Overcome Them
Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, Sabine Süsstrunk
We analyze the influence of adversarial training on the loss landscape of machine learning models.
To this end, we first provide analytical studies of the properties of adversarial loss functions under different adversarial budgets.
We then demonstrate that the adversarial loss landscape is less favorable to optimization, due to increased curvature and more scattered gradients.
Our conclusions are validated by numerical analyses, which show that training under large adversarial budgets impede the escape from suboptimal random initialization, cause non-vanishing gradients and make the models' minima found sharper.
Based on these observations, we show that a periodic adversarial scheduling (PAS) strategy can effectively overcome these challenges, yielding better results than vanilla adversarial training while being much less sensitive to the choice of learning rate.
**********

User-Dependent Neural Sequence Models for Continuous-Time Event Data
Alex Boyd, Robert Bamler, Stephan Mandt, Padhraic Smyth
Continuous-time event data are common in applications such as individual behavior data, financial transactions, and medical health records. Modeling such data can be very challenging, in particular for applications with many different types of events,since it requires a model to predict the event types as well as the time of occurrence. Recurrent neural networks that parameterize time-varying intensity functions are the current state-of-the-art for predictive modeling with such data. These models typically assume that all event sequences come from the same data distribution. However, in many applications event sequences are generated by different sources,or users, and their characteristics can be very different. In this paper, we extend the broad class of neural marked point process models to mixtures of latent embeddings,where each mixture component models the characteristic traits of a given user. Our approach relies on augmenting these models with a latent variable that encodes user characteristics, represented by a mixtu

re model over user behavior that is trained via amortized variational inference. We evaluate our methods on four large real-world datasets and demonstrate systematic improvements from our approach over existing work for a variety of predictive metrics such as log-likelihood, next event ranking, and source-of-sequence identification.
**********

Active Structure Learning of Causal DAGs via Directed Clique Trees
Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, Karthikeyan Shanmugam
A growing body of work has begun to study intervention design for efficient structure learning of causal directed acyclic graphs (DAGs).
A typical setting is a \emph{causally sufficient} setting, i.e. a system with no latent confounders, selection bias, or feedback, when the essential graph of the observational equivalence class (EC) is given as an input and interventions are assumed to be noiseless.
Most existing works focus on \textit{worst-case} or \textit{average-case} lower bounds for the number of interventions required to orient a DAG. These worst-case lower bounds only establish that the largest clique in the essential graph \textit{could} make it difficult to learn the true DAG.
In this work, we develop a \textit{universal} lower bound for single-node interventions that establishes that the largest clique is \textit{always} a fundamental impediment to structure learning.
Specifically, we present a decomposition of a DAG into independently orientable components through \emph{directed clique trees} and use it to prove that
the number of single-node interventions necessary to orient any DAG in an EC is at least the sum of half the size of the largest cliques in each chain component of the essential graph.
Moreover, we present a two-phase intervention design algorithm that, under certain conditions on the chordal skeleton, matches the optimal number of interventions up to a multiplicative logarithmic factor in the number of maximal cliques.
We show via synthetic experiments that our algorithm can scale to much larger graphs than most of the related work and achieves better worst-case performance than other scalable approaches. A code base to recreate these results can be found at \url{https://github.com/csquires/dct-policy}.
**********

Convergence and Stability of Graph Convolutional Networks on Large Random Graphs
Nicolas Keriven, Alberto Bietti, Samuel Vaiter
We study properties of Graph Convolutional Networks (GCNs) by analyzing their behavior on standard models of random graphs, where nodes are represented by random latent variables and edges are drawn according to a similarity kernel. This allows us to overcome the difficulties of dealing with discrete notions such as isomorphisms on very large graphs, by considering instead more natural geometric aspects. We first study the convergence of GCNs to their continuous counterpart as the number of nodes grows. Our results are fully non-asymptotic and are valid for relatively sparse graphs with an average degree that grows logarithmically with the number of nodes. We then analyze the stability of GCNs to small deformations of the random graph model. In contrast to previous studies of stability in discrete settings, our continuous setup allows us to provide more intuitive deformation-based metrics for understanding stability, which have proven useful for explaining the success of convolutional representations on Euclidean domains.
**********

BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization
Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G. Wilson, Eytan Bakshy
Bayesian optimization provides sample-efficient global optimization for a broad range of applications, including automatic machine learning, engineering, physics, and experimental design. We introduce BoTorch, a modern programming framework for Bayesian optimization that combines Monte-Carlo (MC) acquisition functions, a novel sample average approximation optimization approach, auto-differentiation, and variance reduction techniques. BoTorch's modular design facilitates flexi

ble specification and optimization of probabilistic models written in PyTorch, s
implifying implementation of new acquisition functions. Our approach is backed b
y novel theoretical convergence results and made practical by a distinctive algo
rithmic foundation that leverages fast predictive distributions, hardware accele
ration, and deterministic optimization. We also propose a novel "one-shot" formu
lation of the Knowledge Gradient, enabled by a combination of our theoretical an
d software contributions. In experiments, we demonstrate the improved sample eff
iciency of BoTorch relative to other popular libraries.
**********

Reconsidering Generative Objectives For Counterfactual Reasoning
Danni Lu, Chenyang Tao, Junya Chen, Fan Li, Feng Guo, Lawrence Carin
There has been recent interest in exploring generative goals for counterfactual
reasoning, such as individualized treatment effect (ITE) estimation. However, ex
isting solutions often fail to address issues that are unique to causal inferenc
e, such as covariate balancing and (infeasible) counterfactual validation. As a
step towards more flexible, scalable and accurate ITE estimation, we present a n
ovel generative Bayesian estimation framework that integrates representation lea
rning, adversarial matching and causal estimation. By appealing to the Robinson
decomposition, we derive a reformulated variational bound that explicitly target
s the causal effect estimation rather than specific predictive goals. Our proced
ure acknowledges the uncertainties in representation and solves a Fenchel mini-m
ax game to resolve the representation imbalance for better counterfactual genera
lization, justified by new theory. Further, the latent variable formulation empl
oyed enables robustness to unobservable latent confounders, extending the scope
of its applicability. The utility of the proposed solution is demonstrated via a
n extensive set of tests against competing solutions, both under various simulat
ion setups and to real-world datasets, with encouraging results reported.
**********

Robust Federated Learning: The Case of Affine Distribution Shifts
Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, Ali Jadbabaie
Federated learning is a distributed  paradigm that aims at training  models usin
g samples distributed across multiple users in a network while keeping the sampl
es on users’ devices with the aim of efficiency and  protecting users privacy. I
n such settings, the training data is often statistically heterogeneous and mani
fests various distribution shifts across users, which degrades the performance o
f the learnt model. The primary goal of this paper is to develop a robust federa
ted learning algorithm that achieves satisfactory performance against distributi
on shifts in users' samples. To achieve this goal, we first consider a structure
d affine distribution shift in users' data that captures the device-dependent da
ta heterogeneity in federated settings. This perturbation model is applicable to
 various federated learning problems such as image classification where the imag
es undergo device-dependent imperfections, e.g. different intensity, contrast, a
nd brightness. To address affine distribution shifts across users, we propose a
Federated Learning framework Robust to Affine distribution shifts (FLRA) that is
 provably robust against affine Wasserstein shifts to the distribution of observ
ed samples. To solve the FLRA's distributed minimax optimization problem, we pro
pose a fast and efficient optimization method and provide convergence and perfor
mance  guarantees via a gradient Descent Ascent (GDA) method. We further prove g
eneralization error bounds for the learnt classifier to show proper generalizati
on from empirical distribution of samples to the true underlying distribution. W
e perform several numerical experiments to empirically support FLRA. We show tha
t an affine distribution shift indeed suffices to significantly decrease the per
formance of the learnt classifier in a new test user, and our proposed algorithm
 achieves a significant gain in comparison to standard federated learning and ad
versarial training methods.
**********

Quantile Propagation for Wasserstein-Approximate Gaussian Processes
Rui Zhang, Christian Walder, Edwin V. Bonilla, Marian-Andrei Rizoiu, Lexing Xie
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked.  However name changes may cause bibliographic tracking issues.

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

Generating Adjacency-Constrained Subgoals in Hierarchical Reinforcement Learning
Tianren Zhang, Shangqi Guo, Tian Tan, Xiaolin Hu, Feng Chen
Goal-conditioned hierarchical reinforcement learning (HRL) is a promising approa
ch for scaling up reinforcement learning (RL) techniques. However, it often suff
ers from training inefficiency as the action space of the high-level, i.e., the
goal space, is often large. Searching in a large goal space poses difficulties f
or both high-level subgoal generation and low-level policy learning. In this pap
er, we show that this problem can be effectively alleviated by restricting the h
igh-level action space from the whole goal space to a k-step adjacent region of
the current state using an adjacency constraint. We theoretically prove that the
proposed adjacency constraint preserves the optimal hierarchical policy in dete
rministic MDPs, and show that this constraint can be practically implemented by
training an adjacency network that can discriminate between adjacent and non-adj
acent subgoals. Experimental results on discrete and continuous control tasks sh
ow that incorporating the adjacency constraint improves the performance of state
-of-the-art HRL approaches in both deterministic and stochastic environments.
**********

High-contrast “gaudy” images improve the training of deep neural network models
of visual cortex
Benjamin Cowley, Jonathan W. Pillow
A key challenge in understanding the sensory transformations of the visual syste
m is to obtain a highly predictive model that maps natural images to neural resp
onses. Deep neural networks (DNNs) provide a promising candidate for such a mode
l. However, DNNs require orders of magnitude more training data than neuroscient
ists can collect because experimental recording time is severely limited. This m
otivates us to find images that train highly-predictive DNNs with as little trai
ning data as possible. We propose high-contrast, binarized versions of natural i
mages---termed gaudy images---to efficiently train DNNs to predict higher-order
visual cortical responses. In simulation experiments and analyses of real neural
data, we find that training DNNs with gaudy images substantially reduces the nu
mber of training images needed to accurately predict responses to natural images
. We also find that gaudy images, chosen before training, outperform images chos
en during training by active learning algorithms. Thus, gaudy images overemphasi
ze features of natural images that are the most important for efficiently traini
ng DNNs. We believe gaudy images will aid in the modeling of visual cortical neu
rons, potentially opening new scientific questions about visual processing.
**********

Duality-Induced Regularizer for Tensor Factorization Based Knowledge Graph Compl
etion
Zhanqiu Zhang, Jianyu Cai, Jie Wang
Tensor factorization based models have shown great power in knowledge graph comp
letion (KGC). However, their performance usually suffers from the overfitting pr
oblem seriously. This motivates various regularizers---such as the squared Frobe
nius norm and tensor nuclear norm regulariers---while the limited applicability
significantly limits their practical usage. To address this challenge, we propos
e a novel regularizer---namely, \textbf{DU}ality-induced \textbf{R}egul\textbf{A
}rizer (DURA)---which is not only effective in improving the performance of exis
ting models but widely applicable to various methods. The major novelty of DURA
is based on the observation that, for an existing tensor factorization based KGC
model (\textit{primal}), there is often another distance based KGC model (\text
it{dual}) closely associated with it.
**********

Distributed Training with Heterogeneous Data: Bridging Median- and Mean-Based Al
gorithms
Xiangyi Chen, Tiancong Chen, Haoran Sun, Steven Z. Wu, Mingyi Hong
Recently, there is a growing interest in the study of median-based algorithms fo
r distributed non-convex optimization. Two prominent examples include signSGD wi

th majority vote, an effective approach for communication reduction via 1-bit co mpression on the local gradients, and medianSGD, an algorithm recently proposed to ensure robustness against Byzantine workers. The convergence analyses for the se algorithms critically rely on the assumption that all the distributed data ar e drawn iid from the same distribution. However, in applications such as Federat ed Learning, the data across different nodes or machines can be inherently heter ogeneous, which violates such an iid assumption. This work analyzes signSGD and medianSGD in distributed settings with heterogeneous data. We show that these al gorithms are non-convergent whenever there is some disparity between the expecte d median and mean over the local gradients. To overcome this gap, we provide a n ovel gradient correction mechanism that perturbs the local gradients with noise, which we show can provably close the gap between mean and median of the gradien ts. The proposed methods largely preserve nice properties of these median-based algorithms, such as the low per-iteration communication complexity of signSGD, a nd further enjoy global convergence to stationary solutions. Our perturbation te chnique can be of independent interest when one wishes to estimate mean through a median estimator.

**********

H-Mem: Harnessing synaptic plasticity with Hebbian Memory Networks

Thomas Limbacher, Robert Legenstein

The ability to base current computations on memories from the past is critical f or many cognitive tasks such as story understanding. Hebbian-type synaptic plast icity is believed to underlie the retention of memories over medium and long tim e scales in the brain. However, it is unclear how such plasticity processes are integrated with computations in cortical networks. Here, we propose Hebbian Memo ry Networks (H-Mems), a simple neural network model that is built around a core hetero-associative network subject to Hebbian plasticity. We show that the netwo rk can be optimized to utilize the Hebbian plasticity processes for its computat ions. H-Mems can one-shot memorize associations between stimulus pairs and use t hese associations for decisions later on. Furthermore, they can solve demanding question-answering tasks on synthetic stories. Our study shows that neural netw ork models are able to enrich their computations with memories through simple He bbian plasticity processes.

**********

Neural Unsigned Distance Fields for Implicit Function Learning

Julian Chibane, Mohamad Aymen mir, Gerard Pons-Moll

In this work we target a learnable output representation that allows continuous, high resolution outputs of arbitrary shape.

Recent works represent 3D surfaces implicitly with a Neural Network, thereby bre aking previous barriers in resolution, and ability to represent diverse topologi es.

However, neural implicit representations are limited to closed surfaces, which d ivide the space into inside and outside. Many real world objects such as walls o f a scene scanned by a sensor, clothing, or a car with inner structures are not closed.

This constitutes a significant barrier, in terms of data pre-processing (objects need to be artificially closed creating artifacts), and the ability to output o pen surfaces.

In this work, we propose Neural Distance Fields (NDF), a neural network based mo del which predicts the unsigned distance field for arbitrary 3D shapes given spa rse point clouds.

NDF represent surfaces at high resolutions as prior implicit models, but do not require closed surface data, and significantly broaden the class of representabl e shapes in the output.

NDF allow to extract the surface as very dense point clouds and as meshes.

We also show that NDF allow for surface normal calculation and can be rendered u sing a slight modification of sphere tracing.

We find NDF can be used for multi-target regression (multiple outputs for one in put) with techniques that have been exclusively used for rendering in graphics.

Experiments on ShapeNet show that NDF, while simple, is the state-of-the art, an

d allows to reconstruct shapes with inner structures, such as the chairs inside a bus.
Notably, we show that NDF are not restricted to 3D shapes, and can approximate m ore general open surfaces such as curves, manifolds, and functions.
Code is available for research at https://virtualhumans.mpi-inf.mpg.de/ndf/.
**********

Curriculum By Smoothing
Samarth Sinha, Animesh Garg, Hugo Larochelle
Convolutional Neural Networks (CNNs) have shown impressive performance in comput er vision tasks
such as image classification, detection, and segmentation.
Moreover, recent work in Generative Adversarial Networks (GANs) has highlighted the importance of learning
by progressively increasing the difficulty of a learning task Kerras et al.
When learning a network from scratch, the information propagated within the netw ork during the earlier
stages of training can contain distortion artifacts due to noise which can be de trimental to training.
In this paper, we propose an elegant curriculum-based scheme that smoothes the f eature embedding of a CNN
using anti-aliasing or low-pass filters.
We propose to augment the training of CNNs by controlling the amount of high fre quency information
propagated within the CNNs as training progresses, by convolving the output of a CNN feature map
of each layer with a Gaussian kernel.
By decreasing the variance of the Gaussian kernel, we gradually increase the am ount of high-frequency information available within the network for inference.
As the amount of information in the feature maps increases during training,
the network is able to progressively learn better representations of the data.
Our proposed augmented training scheme significantly improves the performance of CNNs on various vision
tasks without either adding additional trainable parameters or an auxiliary regu larization objective.
The generality of our method is demonstrated through empirical performance gains in CNN architectures across four different tasks: transfer learning, cross-task transfer learning,
and generative models.
**********

Fast Transformers with Clustered Attention
Apoorv Vyas, Angelos Katharopoulos, François Fleuret
Transformers have been proven a successful model for a variety of tasks in sequence modeling. However, computing the attention matrix, which is their key component, has quadratic complexity with respect to the sequence length, thus making them prohibitively expensive for large sequences. To address this, we propose clustered attention, which instead of computing the attention for every query, groups queries into clusters and computes attention just for the centroids. To further improve this approximation, we use the computed clusters to identify the keys with the highest attention per query and compute the exact key/query dot products. This results in a model with linear complexity with respect to the sequence length for a fixed number of clusters. We evaluate our approach on two automatic speech recognition datasets and show that our model consistently outperforms vanilla transformers for a given computational budget. Finally, we demonstrate that our model can approximate arbitrarily complex attention distributions with a minimal number of clusters by approximating a pretrained BERT model on GLUE and SQuAD benchmarks with only 25 clusters and no loss in performance.
**********

The Convex Relaxation Barrier, Revisited: Tightened Single-Neuron Relaxations fo r Neural Network Verification

Christian Tjandraatmadja, Ross Anderson, Joey Huchette, Will Ma, KRUNAL KISHOR PATEL, Juan Pablo Vielma

We improve the effectiveness of propagation- and linear-optimization-based neural network verification algorithms with a new tightened convex relaxation for ReLU neurons. Unlike previous single-neuron relaxations which focus only on the univariate input space of the ReLU, our method considers the multivariate input space of the affine pre-activation function preceding the ReLU. Using results from submodularity and convex geometry, we derive an explicit description of the tightest possible convex relaxation when this multivariate input is over a box domain. We show that our convex relaxation is significantly stronger than the commonly used univariate-input relaxation which has been proposed as a natural convex relaxation barrier for verification. While our description of the relaxation may require an exponential number of inequalities, we show that they can be separated in linear time and hence can be efficiently incorporated into optimization algorithms on an as-needed basis. Based on this novel relaxation, we design two polynomial-time algorithms for neural network verification: a linear-programming-based algorithm that leverages the full power of our relaxation, and a fast propagation algorithm that generalizes existing approaches. In both cases, we show that for a modest increase in computational effort, our strengthened relaxation enables us to verify a significantly larger number of instances compared to similar algorithms.

**********

Strongly Incremental Constituency Parsing with Graph Neural Networks

Kaiyu Yang, Jia Deng

Parsing sentences into syntax trees can benefit downstream applications in NLP. Transition-based parsers build trees by executing actions in a state transition system. They are computationally efficient, and can leverage machine learning to predict actions based on partial trees. However, existing transition-based parsers are predominantly based on the shift-reduce transition system, which does not align with how humans are known to parse sentences. Psycholinguistic research suggests that human parsing is strongly incremental—humans grow a single parse tree by adding exactly one token at each step. In this paper, we propose a novel transition system called attach-juxtapose. It is strongly incremental; it represents a partial sentence using a single tree; each action adds exactly one token into the partial tree. Based on our transition system, we develop a strongly incremental parser. At each step, it encodes the partial tree using a graph neural network and predicts an action. We evaluate our parser on Penn Treebank (PTB) and Chinese Treebank (CTB). On PTB, it outperforms existing parsers trained with only constituency trees; and it performs on par with state-of-the-art parsers that use dependency trees as additional training data. On CTB, our parser establishes a new state of the art. Code is available at https://github.com/princeton-vl/attach-juxtapose-parser.

**********

AOT: Appearance Optimal Transport Based Identity Swapping for Forgery Detection

Hao Zhu, Chaoyou Fu, Qianyi Wu, Wayne Wu, Chen Qian, Ran He

Recent studies have shown that the performance of forgery detection can be improved with diverse and challenging Deepfakes datasets. However, due to the lack of Deepfakes datasets with large variance in appearance, which can be hardly produced by recent identity swapping methods, the detection algorithm may fail in this situation. In this work, we provide a new identity swapping algorithm with large differences in appearance for face forgery detection. The appearance gaps mainly arise from the large discrepancies in illuminations and skin colors that widely exist in real-world scenarios. However, due to the difficulties of modeling the complex appearance mapping, it is challenging to transfer fine-grained appearances adaptively while preserving identity traits. This paper formulates appearance mapping as an optimal transport problem and proposes an Appearance Optimal Transport model (AOT) to formulate it in both latent and pixel space. Specifically, a relighting generator is designed to simulate the optimal transport plan. It is solved via minimizing the Wasserstein distance of the learned features in the latent space, enabling better performance and less computation than conventio

nal optimization. To further refine the solution of the optimal transport plan, we develop a segmentation game to minimize the Wasserstein distance in the pixel space. A discriminator is introduced to distinguish the fake parts from a mix of real and fake image patches. Extensive experiments reveal that the superiority of our method when compared with state-of-the-art methods and the ability of our generated data to improve the performance of face forgery detection.
**********

Uncertainty-Aware Learning for Zero-Shot Semantic Segmentation
Ping Hu, Stan Sclaroff, Kate Saenko
Zero-shot semantic segmentation (ZSS) aims to classify pixels of novel classes without training examples available. Recently, most ZSS methods focus on learning the visual-semantic correspondence to transfer knowledge from seen classes to unseen classes at the pixel level. Yet, few works study the adverse effects caused by the noisy and outlying training samples in the seen classes. In this paper, we identify this challenge and address it with a novel framework that learns to discriminate noisy samples based on Bayesian uncertainty estimation. Specifically, we model the network outputs with Gaussian and Laplacian distributions, with the variances accounting for the observation noise as well as the uncertainty of input samples. Learning objectives are then derived with the estimated variances playing as adaptive attenuation for individual samples in training. Consequently, our model learns more attentively from representative samples of seen classes while suffering less from noisy and outlying ones, thus providing better reliability and generalization toward unseen categories. We demonstrate the effectiveness of our framework through comprehensive experiments on multiple challenging benchmarks, and show that our method achieves significant accuracy improvement over previous approaches for large open-set segmentation.
**********

Delta-STN: Efficient Bilevel Optimization for Neural Networks using Structured Response Jacobians
Juhan Bae, Roger B. Grosse
Hyperparameter optimization of neural networks can be elegantly formulated as a bilevel optimization problem. While research on bilevel optimization of neural networks has been dominated by implicit differentiation and unrolling, hypernetworks such as Self-Tuning Networks (STNs) have recently gained traction due to their ability to amortize the optimization of the inner objective. In this paper, we diagnose several subtle pathologies in the training of STNs. Based on these observations, we propose the Delta-STN, an improved hypernetwork architecture which stabilizes training and optimizes hyperparameters much more efficiently than STNs. The key idea is to focus on accurately approximating the best-response Jacobian rather than the full best-response function; we achieve this by reparameterizing the hypernetwork and linearizing the network around the current parameters. We demonstrate empirically that our Delta-STN can tune regularization hyperparameters (e.g. weight decay, dropout, number of cutout holes) with higher accuracy, faster convergence, and improved stability compared to existing approaches.
**********

First-Order Methods for Large-Scale Market Equilibrium Computation
Yuan Gao, Christian Kroer
Market equilibrium is a solution concept with many applications such as digital ad markets, fair division, and resource sharing. For many classes of utility functions, equilibria can be captured by convex programs. We develop simple first-order methods suitable for solving these programs for large-scale markets. We focus on three practically-relevant utility classes: linear, quasilinear, and Leontief utilities. Using structural properties of market equilibria under each utility class, we show that the corresponding convex programs can be reformulated as optimization of a structured smooth convex function over a polyhedral set, for which projected gradient achieves linear convergence. To do so, we utilize recent linear convergence results under weakened strong-convexity conditions, and further refine the relevant constants in existing convergence results. Then, we show that proximal gradient (a generalization of projected gradient) with a practical linesearch scheme achieves linear convergence under the Proximal-PL condition,

a recently developed error bound condition for convex composite problems. For quasilinear utilities, we show that Mirror Descent applied to a new convex program achieves sublinear last-iterate convergence and yields a form of Proportional Response dynamics, an elegant, interpretable algorithm for computing market equilibria originally developed for linear utilities. Numerical experiments show that Proportional Response is highly efficient for computing approximate market equilibria, while projected gradient with linesearch can be much faster when higher-accuracy solutions are needed.

**********

Minimax Optimal Nonparametric Estimation of Heterogeneous Treatment Effects

Zijun Gao, Yanjun Han

A central goal of causal inference is to detect and estimate the treatment effects of a given treatment or intervention on an outcome variable of interest, where a member known as the heterogeneous treatment effect (HTE) is of growing popularity in recent practical applications such as the personalized medicine. In this paper, we model the HTE as a smooth nonparametric difference between two less smooth baseline functions, and determine the tight statistical limits of the nonparametric HTE estimation as a function of the covariate geometry. In particular, a two-stage nearest-neighbor-based estimator throwing away observations with poor matching quality is near minimax optimal. We also establish the tight dependence on the density ratio without the usual assumption that the covariate densities are bounded away from zero, where a key step is to employ a novel maximal inequality which could be of independent interest.

**********

Residual Force Control for Agile Human Behavior Imitation and Extended Motion Synthesis

Ye Yuan, Kris Kitani

Reinforcement learning has shown great promise for synthesizing realistic human behaviors by learning humanoid control policies from motion capture data. However, it is still very challenging to reproduce sophisticated human skills like ballet dance, or to stably imitate long-term human behaviors with complex transitions. The main difficulty lies in the dynamics mismatch between the humanoid model and real humans. That is, motions of real humans may not be physically possible for the humanoid model. To overcome the dynamics mismatch, we propose a novel approach, residual force control (RFC), that augments a humanoid control policy by adding external residual forces into the action space. During training, the RFC-based policy learns to apply residual forces to the humanoid to compensate for the dynamics mismatch and better imitate the reference motion. Experiments on a wide range of dynamic motions demonstrate that our approach outperforms state-of-the-art methods in terms of convergence speed and the quality of learned motions. Notably, we showcase a physics-based virtual character empowered by RFC that can perform highly agile ballet dance moves such as pirouette, arabesque and jeté. Furthermore, we propose a dual-policy control framework, where a kinematic policy and an RFC-based policy work in tandem to synthesize multi-modal infinite-horizon human motions without any task guidance or user input. Our approach is the first humanoid control method that successfully learns from a large-scale human motion dataset (Human3.6M) and generates diverse long-term motions. Code and videos are available at https://www.ye-yuan.com/rfc.

**********

A General Method for Robust Learning from Batches

Ayush Jain, Alon Orlitsky

In many applications, data is collected in batches, some of which may be corrupt or even adversarial. Recent work derived optimal robust algorithms for estimating finite distributions in this setting. We develop a general framework of robust learning from batches, and determine the limits of both distribution estimation, and notably, classification, over arbitrary, including continuous, domains.

**********

Not All Unlabeled Data are Equal: Learning to Weight Data in Semi-supervised Learning

Zhongzheng Ren, Raymond Yeh, Alexander Schwing

Existing semi-supervised learning (SSL) algorithms use a single weight to balanc
e the loss of labeled and unlabeled examples, i.e., all unlabeled examples are e
qually weighted. But not all unlabeled data are equal. In this paper we study ho
w to use a different weight for “every” unlabeled example. Manual tuning of all
those weights -- as done in prior work -- is no longer possible. Instead, we adj
ust those weights via an algorithm based on the influence function, a measure of
 a model's dependency on one training example. To make the approach efficient, w
e propose a fast and effective approximation of the influence function. We demon
strate that this technique outperforms state-of-the-art methods on semi-supervis
ed image and language classification tasks.
**********

Hard Negative Mixing for Contrastive Learning
Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, Diane
 Larlus
Contrastive learning has become a key component of self-supervised learning appr
oaches for computer vision. By learning to embed two augmented versions of the s
ame image close to each other and to push the embeddings of different images apa
rt, one can train highly transferable visual representations. As revealed by rec
ent studies, heavy data augmentation and large sets of negatives are both crucia
l in learning such representations. At the same time, data mixing strategies, ei
ther at the image or the feature level, improve both supervised and semi-supervi
sed learning by synthesizing novel examples, forcing networks to learn more robu
st features. In this paper, we argue that an important aspect of contrastive lea
rning, i.e. the effect of hard negatives, has so far been neglected. To get more
 meaningful negative samples, current top contrastive self-supervised learning a
pproaches either substantially increase the batch sizes, or keep very large memo
ry banks; increasing memory requirements, however, leads to diminishing returns
in terms of performance. We therefore start by delving deeper into a top-perform
ing framework and show evidence that harder negatives are needed to facilitate b
etter and faster learning. Based on these observations, and motivated by the suc
cess of data mixing, we propose hard negative mixing strategies at the feature l
evel, that can be computed on-the-fly with a minimal computational overhead. We
exhaustively ablate our approach on linear classification, object detection, and
 instance segmentation and show that employing our hard negative mixing procedur
e improves the quality of visual representations learned by a state-of-the-art s
elf-supervised learning method.
**********

MOReL: Model-Based Offline Reinforcement Learning
Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, Thorsten Joachims
In offline reinforcement learning (RL), the goal is to learn a highly rewarding
policy based solely on a dataset of historical interactions with the environment
. This serves as an extreme test for an agent's ability to effectively use histo
rical data which is known to be critical for efficient RL. Prior work in offline
 RL has been confined almost exclusively to model-free RL approaches. In this wo
rk, we present MOReL, an algorithmic framework for model-based offline RL. This
framework consists of two steps: (a) learning a pessimistic MDP using the offlin
e dataset; (b) learning a near-optimal policy in this pessimistic MDP. The desig
n of the pessimistic MDP is such that for any policy, the performance in the rea
l environment is approximately lower-bounded by the performance in the pessimist
ic MDP. This enables the pessimistic MDP to serve as a good surrogate for purpos
es of policy evaluation and learning. Theoretically, we show that MOReL is minim
ax optimal (up to log factors) for offline RL. Empirically, MOReL matches or exc
eeds state-of-the-art results on widely used offline RL benchmarks. Overall, the
 modular design of MOReL enables translating advances in its components (for e.g
., in model learning, planning etc.) to improvements in offline RL.
**********

Weisfeiler and Leman go sparse: Towards scalable higher-order graph embeddings
Christopher Morris, Gaurav Rattan, Petra Mutzel
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Adversarial Crowdsourcing Through Robust Rank-One Matrix Completion
Qianqian Ma, Alex Olshevsky
We consider the problem of reconstructing a rank-one matrix from a revealed subs
et of its entries when some of the revealed entries are corrupted with perturbat
ions that are unknown and can be arbitrarily large. It is not known which revea
led entries are corrupted. We propose a new algorithm combining alternating min
imization with extreme-value filtering and provide sufficient and necessary cond
itions to recover the original rank-one matrix. In particular, we show that our
proposed algorithm is optimal when the set of revealed entries is given by an E
rd\H os-R\'enyi random graph.
**********

Learning Semantic-aware Normalization for Generative Adversarial Networks
Heliang Zheng, Jianlong Fu, Yanhong Zeng, Jiebo Luo, Zheng-Jun Zha
The recent advances in image generation have been achieved by style-based image
generators. Such approaches learn to disentangle latent factors in different ima
ge scales and encode latent factors as “style” to control image synthesis. Howev
er, existing approaches cannot further disentangle fine-grained semantics from e
ach other, which are often conveyed from feature channels. In this paper, we pro
pose a novel image synthesis approach by learning Semantic-aware relative import
ance for feature channels in Generative Adversarial Networks (SariGAN). Such a m
odel disentangles latent factors according to the semantic of feature channels b
y channel-/group- wise fusion of latent codes and feature channels. Particularl
y, we learn to cluster feature channels by semantics and propose an adaptive gro
up-wise Normalization (AdaGN) to independently control the styles of different c
hannel groups. For example, we can adjust the statistics of channel groups for a
human face to control the open and close of the mouth, while keeping other faci
al features unchanged. We propose to use adversarial training, a channel groupin
g loss, and a mutual information loss for joint optimization, which not only ena
bles high-fidelity image synthesis but leads to superior interpretable propertie
s. Extensive experiments show that our approach outperforms the SOTA style-based
approaches in both unconditional image generation and conditional image inpaint
ing tasks.
**********

Differentiable Causal Discovery from Interventional Data
Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Juli
en, Alexandre Drouin
Learning a causal directed acyclic graph from data is a challenging task that in
volves solving a combinatorial problem for which the solution is not always iden
tifiable. A new line of work reformulates this problem as a continuous constrain
ed optimization one, which is solved via the augmented Lagrangian method. Howeve
r, most methods based on this idea do not make use of interventional data, which
can significantly alleviate identifiability issues. This work constitutes a new
step in this direction by proposing a theoretically-grounded method based on ne
ural networks that can leverage interventional data. We illustrate the flexibili
ty of the continuous-constrained framework by taking advantage of expressive neu
ral architectures such as normalizing flows. We show that our approach compares
favorably to the state of the art in a variety of settings, including perfect an
d imperfect interventions for which the targeted nodes may even be unknown.
**********

One-sample Guided Object Representation Disassembling
Zunlei Feng, Yongming He, Xinchao Wang, Xin Gao, Jie Lei, Cheng Jin, Mingli Song
The ability to disassemble the features of objects and background is crucial for
many machine learning tasks, including image classification, image editing, vis
ual concepts learning, and so on. However, existing (semi-)supervised methods al
l need a large amount of annotated samples, while unsupervised methods can't han
dle real-world images with complicated backgrounds. In this paper, we introduce
the One-sample Guided Object Representation Disassembling (One-GORD) method, whi

ch only requires one annotated sample for each object category to learn disassembled object representation from unannotated images. For the annotated one-sample, we first adopt some data augmentation strategies to generate some synthetic samples, which can guide the disassembling of the object features and background features. For the unannotated images, two self-supervised mechanisms: dual-swapping and fuzzy classification are introduced to disassemble object features from the background with the guidance of annotated one-sample. What's more, we devise two metrics to evaluate the disassembling performance from the perspective of representation and image, respectively. Experiments demonstrate that the One-GORD achieves competitive dissembling performance and can handle natural scenes with complicated backgrounds.

**********

Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate

Akifumi Okuno, Hidetoshi Shimodaira

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

Siddharth Vishwanath, Kenji Fukumizu, Satoshi Kuriki, Bharath K. Sriperumbudur

Persistent homology has become an important tool for extracting geometric and topological features from data, whose multi-scale features are summarized in a persistence diagram. From a statistical perspective, however, persistence diagrams are very sensitive to perturbations in the input space. In this work, we develop a framework for constructing robust persistence diagrams from superlevel filtrations of robust density estimators constructed using reproducing kernels. Using an analogue of the influence function on the space of persistence diagrams, we establish the proposed framework to be less sensitive to outliers. The robust persistence diagrams are shown to be consistent estimators in the bottleneck distance, with the convergence rate controlled by the smoothness of the kernel — this, in turn, allows us to construct uniform confidence bands in the space of persistence diagrams. Finally, we demonstrate the superiority of the proposed approach on benchmark datasets.

**********

Contextual Games: Multi-Agent Learning with Side Information

Pier Giuseppe Sessa, Ilija Bogunovic, Andreas Krause, Maryam Kamgarpour

We formulate the novel class of contextual games, a type of repeated games driven by contextual information at each round. By means of kernel-based regularity assumptions, we model the correlation between different contexts and game outcomes and propose a novel online (meta) algorithm that exploits such correlations to minimize the contextual regret of individual players. We define game-theoretic notions of contextual Coarse Correlated Equilibria (c-CCE) and optimal contextual welfare for this new class of games and show that c-CCEs and optimal welfare can be approached whenever players' contextual regrets vanish. Finally, we empirically validate our results in a traffic routing experiment, where our algorithm leads to better performance and higher welfare compared to baselines that do not exploit the available contextual information or the correlations present in the game.

**********

Goal-directed Generation of Discrete Structures with Conditional Generative Models

Amina Mollaysa, Brooks Paige, Alexandros Kalousis

Despite recent advances, goal-directed generation of structured discrete data remains challenging. For problems such as program synthesis (generating source code) and materials design (generating molecules), finding examples which satisfy desired constraints or exhibit desired properties is difficult. In practice, expensive heuristic search or reinforcement learning algorithms are often employed. In this paper, we investigate the use of conditional generative models which dir

ectly attack this inverse problem, by modeling the distribution of discrete stru
ctures given
properties of interest. Unfortunately, the maximum likelihood training of such m
odels often fails with the samples from the generative model inadequately respec
ting the input properties. To address this, we introduce a novel approach to dir
ectly optimize a reinforcement learning objective, maximizing an expected reward
. We avoid high-variance score-function estimators that would otherwise be requi
red by sampling from an approximation to the normalized rewards, allowing simple
 Monte Carlo estimation of model gradients. We test our methodology on two tasks
: generating molecules with user-defined properties and identifying short python
 expressions which evaluate to a given target value. In both cases, we find impr
ovements over maximum likelihood estimation and other baselines.
**********

Beyond Lazy Training for Over-parameterized Tensor Decomposition
Xiang Wang, Chenwei Wu, Jason D. Lee, Tengyu Ma, Rong Ge
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Denoised Smoothing: A Provable Defense for Pretrained Classifiers
Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, 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-auth
ors prior to requesting a name change in the electronic proceedings.
**********

Minibatch Stochastic Approximate Proximal Point Methods
Hilal Asi, Karan Chadha, Gary Cheng, John C. Duchi
We extend the Approximate-Proximal Point (aProx) family of model-based methods f
or solving stochastic convex optimization problems, including stochastic subgrad
ient, proximal point, and bundle methods, to the minibatch setting. To do this,
we propose two minibatched algorithms for which we prove a non-asymptotic upper
bound on the rate of convergence, revealing a linear speedup in minibatch size.
In contrast to standard stochastic gradient methods, these methods may have line
ar speedup in the minibatch setting even for non-smooth functions. Our algorithm
s maintain the desirable traits characteristic of the aProx family, such as robu
stness to initial step size choice. Additionally, we show improved convergence r
ates for "interpolation" problems, which (for example) gives a new parallelizati
on strategy for alternating projections. We corroborate our theoretical results
with extensive empirical testing, which demonstrates the gains provided by accur
ate modeling and minibatching.
**********

Attribute Prototype Network for Zero-Shot Learning
Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, Zeynep Akata
From the beginning of zero-shot learning research, visual attributes have been s
hown to play an important role. In order to better transfer attribute-based know
ledge from known to unknown classes, we argue that an image representation with
integrated attribute localization ability would be beneficial for zero-shot lear
ning. To this end, we propose a novel zero-shot representation learning framewor
k that jointly learns discriminative global and local features using only class-
level attributes. While a visual-semantic embedding layer learns global features
, local features are learned through an attribute prototype network that simulta
neously regresses and decorrelates attributes from intermediate features. We sho
w that our locality augmented image representations achieve a new state-of-the-a
rt on three zero-shot learning benchmarks. As an additional benefit, our model p
oints to the visual evidence of the attributes in an image, e.g. for the CUB dat
aset, confirming the improved attribute localization ability of our image repres
entation.
**********

CrossTransformers: spatially-aware few-shot transfer
Carl Doersch, Ankush Gupta, Andrew Zisserman
Given new tasks with very little data---such as new classes in a classification problem or a domain shift in the input---performance of modern vision systems degrades remarkably quickly. In this work, we illustrate how the neural network representations which underpin modern vision systems are subject to supervision collapse, whereby they lose any information that is not necessary for performing the training task, including information that may be necessary for transfer to new tasks or domains. We then propose two methods to mitigate this problem. First, we employ self-supervised learning to encourage general-purpose features that transfer better. Second, we propose a novel Transformer based neural network architecture called CrossTransformers, which can take a small number of labeled images and an unlabeled query, find coarse spatial correspondence between the query and the labeled images, and then infer class membership by computing distances between spatially-corresponding features. The result is a classifier that is more robust to task and domain shift, which we demonstrate via state-of-the-art performance on Meta-Dataset, a recent dataset for evaluating transfer from ImageNet to many other vision datasets.
**********
Learning Latent Space Energy-Based Prior Model
Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, Ying Nian Wu
We propose an energy-based model (EBM) in the latent space of a generator model, so that the EBM serves as a prior model that stands on the top-down network of the generator model. Both the latent space EBM and the top-down network can be learned jointly by maximum likelihood, which involves short-run MCMC sampling from both the prior and posterior distributions of the latent vector. Due to the low dimensionality of the latent space and the expressiveness of the top-down network, a simple EBM in latent space can capture regularities in the data effectively, and MCMC sampling in latent space is efficient and mixes well. We show that the learned model exhibits strong performances in terms of image and text generation and anomaly detection. The one-page code can be found in supplementary materials.
**********
SEVIR : A Storm Event Imagery Dataset for Deep Learning Applications in Radar and Satellite Meteorology
Mark Veillette, Siddharth Samsi, Chris Mattioli
Modern deep learning approaches have shown promising results in meteorological applications like precipitation nowcasting, synthetic radar generation, front detection and several others.  In order to effectively train and validate these complex algorithms, large and diverse datasets containing high-resolution imagery are required. Petabytes of weather data, such as from the Geostationary Environmental Satellite System (GOES) and the Next-Generation Radar (NEXRAD) system, are available to the public; however, the size and complexity of these datasets is a hindrance to developing and training deep models. To help address this problem, we introduce the Storm EVent ImagRy (SEVIR) dataset - a single, rich dataset that combines spatially and temporally aligned data from multiple sensors, along with baseline implementations of deep learning models and evaluation metrics, to accelerate new algorithmic innovations. SEVIR is an annotated, curated and spatio-temporally aligned dataset containing over 10,000 weather events that each consist of 384 km x 384 km image sequences spanning 4 hours of time. Images in SEVIR were sampled and aligned across five different data types: three channels (C02, C09, C13) from the GOES-16 advanced baseline imager, NEXRAD vertically integrated liquid mosaics, and GOES-16 Geostationary Lightning Mapper (GLM) flashes.  Many events in SEVIR were selected and matched to the NOAA Storm Events database so that additional descriptive information such as storm impacts and storm descriptions can be linked to the rich imagery provided by the sensors. We describe the data collection methodology and illustrate the applications of this dataset with two examples of deep learning in meteorology: precipitation nowcasting and synthetic weather radar generation. In addition, we also describe a set of metrics that can be used to evaluate the outputs of these models. The SEVIR dataset an

d baseline implementations of selected applications are available for download.
**********
Lightweight Generative Adversarial Networks for Text-Guided Image Manipulation
Bowen Li, Xiaojuan Qi, Philip Torr, Thomas Lukasiewicz
We propose a novel lightweight generative adversarial network for efficient imag e manipulation using natural language descriptions. To achieve this, a new word-level discriminator is proposed, which provides the generator with fine-grained training feedback at word-level, to facilitate training a lightweight generator that has a small number of parameters, but can still correctly focus on specific visual attributes of an image, and then edit them without affecting other conte nts that are not described in the text. Furthermore, thanks to the explicit trai ning signal related to each word, the discriminator can also be simplified to ha ve a lightweight structure. Compared with the state of the art, our method has a much smaller number of parameters, but still achieves a competitive manipulatio n performance. Extensive experimental results demonstrate that our method can be tter disentangle different visual attributes, then correctly map them to corresp onding semantic words, and thus achieve a more accurate image modification using natural language descriptions.
**********
High-Dimensional Contextual Policy Search with Unknown Context Rewards using Bay esian Optimization
Qing Feng , Ben Letham, Hongzi Mao, Eytan Bakshy
Contextual policies are used in many settings to customize system parameters and actions to the specifics of a particular setting. In some real-world settings, such as randomized controlled trials or A/B tests, it may not be possible to mea sure policy outcomes at the level of context—we observe only aggregate rewards a cross a distribution of contexts. This makes policy optimization much more diffi cult because we must solve a high-dimensional optimization problem over the enti re space of contextual policies, for which existing optimization methods are not suitable. We develop effective models that leverage the structure of the search space to enable contextual policy optimization directly from the aggregate rewa rds using Bayesian optimization. We use a collection of simulation studies to ch aracterize the performance and robustness of the models, and show that our appro ach of inferring a low-dimensional context embedding performs best. Finally, we show successful contextual policy optimization in a real-world video bitrate pol icy problem.
**********
Model Fusion via Optimal Transport
Sidak Pal Singh, Martin Jaggi
Combining different models is a widely used paradigm in machine learning applica tions. While the most common approach is to form an ensemble of models and avera ge their individual predictions, this approach is often rendered infeasible by g iven resource constraints in terms of memory and computation, which grow linearl y with the number of models. We present a layer-wise model fusion algorithm for neural networks that utilizes optimal transport to (soft-) align neurons across the models before averaging their associated parameters.
**********
On the Stability and Convergence of Robust Adversarial Reinforcement Learning: A Case Study on Linear Quadratic Systems
Kaiqing Zhang, Bin Hu, Tamer Basar
Reinforcement learning (RL) algorithms can fail to generalize due to the gap bet ween the simulation and the real world. One standard remedy is to use robust adv ersarial RL (RARL) that accounts for this gap during the policy training, by mod eling the gap as an adversary against the training agent. In this work, we reexa mine the effectiveness of RARL under a fundamental robust control setting: the l inear quadratic (LQ) case. We first observe that the popular RARL scheme that gr eedily alternates agents' updates can easily destabilize the system. Motivated b y this, we propose several other policy-based RARL algorithms whose convergence behaviors are then studied both empirically and theoretically. We find: i) the c onventional RARL framework (Pinto et al., 2017) can learn a destabilizing policy

if the initial policy does not enjoy the robust stability property against the adversary; and ii) with robustly stabilizing initializations, our proposed double-loop RARL algorithm provably converges to the global optimal cost while maintaining robust stability on-the-fly. We also examine the stability and convergence issues of other variants of policy-based RARL algorithms, and then discuss several ways to learn robustly stabilizing initializations. From a robust control perspective, we aim to provide some new and critical angles about RARL, by identifying and addressing the stability issues in this fundamental LQ setting in continuous control. Our results make an initial attempt toward better theoretical understandings of policy-based RARL, the core approach in Pinto et al., 2017.
**********

Learning Individually Inferred Communication for Multi-Agent Cooperation
Ziluo Ding, Tiejun Huang, Zongqing Lu
Communication lays the foundation for human cooperation. It is also crucial for multi-agent cooperation. However, existing work focuses on broadcast communication, which is not only impractical but also leads to information redundancy that could even impair the learning process. To tackle these difficulties, we propose Individually Inferred Communication (I2C), a simple yet effective model to enable agents to learn a prior for agent-agent communication. The prior knowledge is learned via causal inference and realized by a feed-forward neural network that maps the agent's local observation to a belief about who to communicate with. The influence of one agent on another is inferred via the joint action-value function in multi-agent reinforcement learning and quantified to label the necessity of agent-agent communication. Furthermore, the agent policy is regularized to better exploit communicated messages. Empirically, we show that I2C can not only reduce communication overhead but also improve the performance in a variety of multi-agent cooperative scenarios, comparing to existing methods.
**********

Set2Graph: Learning Graphs From Sets
Hadar Serviansky, Nimrod Segol, Jonathan Shlomi, Kyle Cranmer, Eilam Gross, Haggai Maron, Yaron Lipman
Many problems in machine learning (ML) can be cast as learning functions from sets to graphs, or more generally to hypergraphs; in short, Set2Graph functions. Examples include clustering, learning vertex and edge features on graphs, and learning features on triplets in a collection.
**********

Graph Random Neural Networks for Semi-Supervised Learning on Graphs
Wenzheng Feng, Jie Zhang, Yuxiao Dong, Yu Han, Huanbo Luan, Qian Xu, Qiang Yang, Evgeny Kharlamov, Jie Tang
We study the problem of semi-supervised learning on graphs, for which graph neural networks (GNNs) have been extensively explored. However, most existing GNNs inherently suffer from the limitations of over-smoothing, non-robustness, and weak-generalization when labeled nodes are scarce. In this paper, we propose a simple yet effective framework—GRAPH RANDOM NEURAL NETWORKS (GRAND)—to address these issues. In GRAND, we first design a random propagation strategy to perform graph data augmentation. Then we leverage consistency regularization to optimize the prediction consistency of unlabeled nodes across different data augmentations. Extensive experiments on graph benchmark datasets suggest that GRAND significantly outperforms state-of- the-art GNN baselines on semi-supervised node classification. Finally, we show that GRAND mitigates the issues of over-smoothing and non-robustness, exhibiting better generalization behavior than existing GNNs. The source code of GRAND is publicly available at https://github.com/Grand20/grand.
**********

Gradient Boosted Normalizing Flows
Robert Giaquinto, Arindam Banerjee
By chaining a sequence of differentiable invertible transformations, normalizing flows (NF) provide an expressive method of posterior approximation, exact density evaluation, and sampling. The trend in normalizing flow literature has been to devise deeper, more complex transformations to achieve greater flexibility. We propose an alternative: Gradient Boosted Normalizing Flows (GBNF) model a densi

ty by successively adding new NF components with gradient boosting. Under the bo
osting framework, each new NF component optimizes a weighted likelihood objectiv
e, resulting in new components that are fit to the suitable residuals of the pre
viously trained components. The GBNF formulation results in a mixture model stru
cture, whose flexibility increases as more components are added. Moreover, GBNFs
 offer a wider, as opposed to strictly deeper, approach that improves existing N
Fs at the cost of additional training---not more complex transformations. We dem
onstrate the effectiveness of this technique for density estimation and, by coup
ling GBNF with a variational autoencoder, generative modeling of images. Our res
ults show that GBNFs outperform their non-boosted analog, and, in some cases, pr
oduce better results with smaller, simpler flows.
**********

Open Graph Benchmark: Datasets for Machine Learning on Graphs
Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Mic
hele Catasta, Jure Leskovec
We present the Open Graph Benchmark (OGB), a diverse set of challenging and real
istic benchmark datasets to facilitate scalable, robust, and reproducible graph
machine learning (ML) research. OGB datasets are large-scale, encompass multiple
 important graph ML tasks, and cover a diverse range of domains, ranging from so
cial and information networks to biological networks, molecular graphs, source c
ode ASTs, and knowledge graphs. For each dataset, we provide a unified evaluatio
n protocol using meaningful application-specific data splits and evaluation metr
ics. In addition to building the datasets, we also perform extensive benchmark e
xperiments for each dataset. Our experiments suggest that OGB datasets present s
ignificant challenges of scalability to large-scale graphs and out-of-distributi
on generalization under realistic data splits, indicating fruitful opportunities
 for future research. Finally, OGB provides an automated end-to-end graph ML pip
eline that simplifies and standardizes the process of graph data loading, experi
mental setup, and model evaluation. OGB will be regularly updated and welcomes i
nputs from the community. OGB datasets as well as data loaders, evaluation scrip
ts, baseline code, and leaderboards are publicly available at https://ogb.stanfo
rd.edu .
**********

Towards Understanding Hierarchical Learning: Benefits of Neural Representations
Minshuo Chen, Yu Bai, Jason D. Lee, Tuo Zhao, Huan Wang, Caiming Xiong, Richard
Socher
Requests for name changes in the electronic proceedings will be accepted with 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.
**********

Texture Interpolation for Probing Visual Perception
Jonathan Vacher, Aida Davila, Adam Kohn, Ruben  Coen-Cagli
Texture synthesis models are important tools for understanding visual processing
. In particular, statistical approaches based on neurally relevant features have
 been instrumental in understanding aspects of visual perception and of neural c
oding. New deep learning-based approaches further improve the quality of synthet
ic textures. Yet, it is still unclear why deep texture synthesis performs so wel
l, and applications of this new framework to probe visual perception are scarce.
 Here, we show that distributions of deep convolutional neural network (CNN) act
ivations of a texture are well described by elliptical distributions and therefo
re, following optimal transport theory, constraining their mean and covariance i
s sufficient to generate new texture samples. Then, we propose the natural geode
sics (ie the shortest path between two points) arising with the optimal transpor
t metric to interpolate between arbitrary textures. Compared to other CNN-based
approaches, our interpolation method appears to match more closely the geometry
of texture perception, and our mathematical framework is better suited to study
its statistical nature. We apply our method by measuring the perceptual scale as
sociated to the interpolation parameter in human observers, and the neural sensi
tivity of different areas of visual cortex in macaque monkeys.

**********
Hierarchical Neural Architecture Search for Deep Stereo Matching
Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, Zongyuan Ge
To reduce the human efforts in neural network design, Neural Architecture Search (NAS) has been applied with remarkable success to various high-level vision tasks such as classification and semantic segmentation. The underlying idea for the NAS algorithm is straightforward, namely, to allow the network the ability to choose among a set of operations (\eg convolution with different filter sizes), one is able to find an optimal architecture that is better adapted to the problem at hand. However, so far the success of NAS has not been enjoyed by low-level geometric vision tasks such as stereo matching. This is partly due to the fact that state-of-the-art deep stereo matching networks, designed by humans, are already sheer in size. Directly applying the NAS to such massive structures is computationally prohibitive based on the currently available mainstream computing resources.
In this paper, we propose the first \emph{end-to-end} hierarchical NAS framework for deep stereo matching by incorporating task-specific human knowledge into the neural architecture search framework. Specifically, following the gold standard pipeline for deep stereo matching (\ie, feature extraction -- feature volume construction and dense matching), we optimize the architectures of the entire pipeline jointly. Extensive experiments show that our searched network outperforms all state-of-the-art deep stereo matching architectures and is ranked at the top 1 accuracy on KITTI stereo 2012, 2015, and Middlebury benchmarks, as well as the top 1 on SceneFlow dataset with a substantial improvement on the size of the network and the speed of inference. Code available at https://github.com/XuelianCheng/LEAStereo.
**********
MuSCLE: Multi Sweep Compression of LiDAR using Deep Entropy Models
Sourav Biswas, Jerry Liu, Kelvin Wong, Shenlong Wang, Raquel Urtasun
We present a novel compression algorithm for reducing the storage of LiDAR sensory data streams. Our model exploits spatio-temporal relationships across multiple LIDAR sweeps to reduce the bitrate of both geometry and intensity values. Towards this goal, we propose a novel conditional entropy model that models the probabilities of the octree symbols, by considering both coarse level geometry and previous sweeps' geometric and intensity information. We then exploit the learned probability to encode the full data-stream into a compact one. Our experiments demonstrate that our method significantly reduces the joint geometry and intensity bitrate over prior state-of-the-art LiDAR compression methods, with a reduction of 7–17% and 15–35% on the UrbanCity and SemanticKITTI datasets respectively.
**********
Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy
Edward Moroshko, Blake E. Woodworth, Suriya Gunasekar, Jason D. Lee, Nati Srebro, Daniel Soudry
We provide a detailed asymptotic study of gradient flow trajectories and their implicit optimization bias when minimizing the exponential loss over "diagonal linear networks". This is the simplest model displaying a transition between "kernel" and non-kernel ("rich" or "active") regimes. We show how the transition is controlled by the relationship between the initialization scale and how accurately we minimize the training loss. Our results indicate that some limit behavior of gradient descent only kick in at ridiculous training accuracies (well beyond 10^-100). Moreover, the implicit bias at reasonable initialization scales and training accuracies is more complex and not captured by these limits.
**********
Focus of Attention Improves Information Transfer in Visual Features
Matteo Tiezzi, Stefano Melacci, Alessandro Betti, Marco Maggini, Marco Gori
Unsupervised learning from continuous visual streams is a challenging problem that cannot be naturally and efficiently managed in the classic batch-mode setting of computation. The information stream must be carefully processed accordingly

to an appropriate spatio-temporal distribution of the visual data, while most approaches of learning commonly assume uniform probability density. In this paper we focus on unsupervised learning for transferring visual information in a truly online setting by using a computational model that is inspired to the principle of least action in physics. The maximization of the mutual information is carried out by a temporal process which yields online estimation of the entropy terms. The model, which is based on second-order differential equations, maximizes the information transfer from the input to a discrete space of symbols related to the visual features of the input, whose computation is supported by hidden neurons. In order to better structure the input probability distribution, we use a human-like focus of attention model that, coherently with the information maximization model, is also based on second-order differential equations. We provide experimental results to support the theory by showing that the spatio-temporal filtering induced by the focus of attention allows the system to globally transfer more information from the input stream over the focused areas and, in some contexts, over the whole frames with respect to the unfiltered case that yields uniform probability distributions.

**********

Auditing Differentially Private Machine Learning: How Private is Private SGD?
Matthew Jagielski, Jonathan Ullman, Alina Oprea

We investigate whether Differentially Private SGD offers better privacy in practice than what is guaranteed by its state-of-the-art analysis. We do so via novel data poisoning attacks, which we show correspond to realistic privacy attacks. While previous work (Ma et al., arXiv 2019) proposed this connection between differential privacy and data poisoning as a defense against data poisoning, our use as a tool for understanding the privacy of a specific mechanism is new. More generally, our work takes a quantitative, empirical approach to understanding the privacy afforded by specific implementations of differentially private algorithms that we believe has the potential to complement and influence analytical work on differential privacy.

**********

A Dynamical Central Limit Theorem for Shallow Neural Networks
Zhengdao Chen, Grant Rotskoff, Joan Bruna, Eric Vanden-Eijnden

Recent theoretical work has characterized the dynamics and convergence properties for wide shallow neural networks trained via gradient descent; the asymptotic regime in which the number of parameters tends towards infinity has been dubbed the "mean-field" limit. At initialization, the randomly sampled parameters lead to a deviation from the mean-field limit that is dictated by the classical central limit theorem (CLT). However, the dynamics of training introduces correlations among the parameters raising the question of how the fluctuations evolve during training. Here, we analyze the mean-field dynamics as a Wasserstein gradient flow and prove that the deviations from the mean-field evolution scaled by the width, in the width-asymptotic limit, remain bounded throughout training. This observation has implications for both the approximation rate and the generalization: the upper bound we obtain is controlled by a Monte-Carlo type resampling error, which importantly does not depend on dimension. We also relate the bound on the fluctuations to the total variation norm of the measure to which the dynamics converges, which in turn controls the generalization error.

**********

Measuring Systematic Generalization in Neural Proof Generation with Transformers
Nicolas Gontier, Koustuv Sinha, Siva Reddy, Chris Pal

We are interested in understanding how well Transformer language models (TLMs) can perform reasoning tasks when trained on knowledge encoded in the form of natural language.

We investigate their systematic generalization abilities on a logical reasoning task in natural language, which involves reasoning over relationships between entities grounded in first-order logical proofs.

Specifically, we perform soft theorem-proving by leveraging TLMs to generate natural language proofs.

We test the generated proofs for logical consistency, along with the accuracy of

the final inference.
We observe length-generalization issues when evaluated on longer-than-trained sequences.
However, we observe TLMs improve their generalization performance after being exposed to longer, exhaustive proofs.
In addition, we discover that TLMs are able to generalize better using backward-chaining proofs compared to their forward-chaining counterparts, while they find it easier to generate forward chaining proofs.
We observe that models that are not trained to generate proofs are better at generalizing to problems based on longer proofs.
This suggests that Transformers have efficient internal reasoning strategies that are harder to interpret.
These results highlight the systematic generalization behavior of TLMs in the context of logical reasoning, and we believe this work motivates deeper inspection of their underlying reasoning strategies.
**********

Big Self-Supervised Models are Strong Semi-Supervised Learners
Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, Geoffrey E. Hinton
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. 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 from Label Proportions: A Mutual Contamination Framework
Clayton Scott, Jianxin Zhang
Learning from label proportions (LLP) is a weakly supervised setting for classification in which unlabeled training instances are grouped into bags, and each bag is annotated with the proportion of each class occurring in that bag. Prior work on LLP has yet to establish a consistent learning procedure, nor does there exist a theoretically justified, general purpose training criterion. In this work we address these two issues by posing LLP in terms of mutual contamination models (MCMs), which have recently been applied successfully to study various other weak supervision settings. In the process, we establish several novel technical results for MCMs, including unbiased losses and generalization error bounds under non-iid sampling plans. We also point out the limitations of a common experimental setting for LLP, and propose a new one based on our MCM framework.
**********

Fast Matrix Square Roots with Applications to Gaussian Processes and Bayesian Optimization
Geoff Pleiss, Martin Jankowiak, David Eriksson, Anil Damle, Jacob Gardner
Matrix square roots and their inverses arise frequently in machine learning, e.g., when sampling from high-dimensional Gaussians N(0,K) or “whitening” a vector b against covariance matrix K. While existing methods typically require O(N^3) computation, we introduce a highly-efficient quadratic-time algorithm for computing K^{1/2}b, K^{-1/2}b, and their derivatives through matrix-vector multiplication (MVMs). Our method combines Krylov subspace methods with a rational approximation and typically achieves 4 decimal places of accuracy with fewer than 100 MVMs. Moreover, the backward pass requires little additional computation. We demonstrate our method's applicability on matrices as large as 50,000 by 50,000 - well beyond traditional methods - with little approximation error. Applying this increased scalability to variational Gaussian processes, Bayesian optimization, and Gibbs sampling results in more powerful models with higher accuracy. In particular, we perform variational GP inference with up to 10,000 inducing points and perform Gibbs sampling on a 25,000-dimensional problem.
**********

Self-Adaptively Learning to Demoiré from Focused and Defocused Image Pairs
Lin Liu, Shanxin Yuan, Jianzhuang Liu, Liping Bao, Gregory Slabaugh, Qi Tian
Moiré artifacts are common in digital photography, resulting from the interference between high-frequency scene content and the color filter array of the camera. Existing deep learning-based demoiréing methods trained on large scale dataset

s are limited in handling various complex moiré patterns, and mainly focus on de moiréing of photos taken of digital displays. Moreover, obtaining moiré-free gro und-truth in natural scenes is difficult but needed for training. In this paper, we propose a self-adaptive learning method for demoiréing a high-frequency imag e, with the help of an additional defocused moiré-free blur image. Given an imag e degraded with moiré artifacts and a moiré-free blur image, our network predict s a moiré-free clean image and a blur kernel with a self-adaptive strategy that does not require an explicit training stage, instead performing test-time adapta tion. Our model has two sub-networks and works iteratively. During each iteratio n, one sub-network takes the moiré image as input, removing moiré patterns and r estoring image details, and the other sub-network estimates the blur kernel from the blur image. The two sub-networks are jointly optimized. Extensive experimen ts demonstrate that our method outperforms state-of-the-art methods and can prod uce high-quality demoiréd results. It can generalize well to the task of removin g moiré artifacts caused by display screens. In addition, we build a new moiré d ataset, including images with screen and texture moiré artifacts. As far as we k now, this is the first dataset with real texture moiré patterns.
**********

Confounding-Robust Policy Evaluation in Infinite-Horizon Reinforcement Learning
Nathan Kallus, Angela Zhou
Off-policy evaluation of sequential decision policies from observational data is necessary in applications of batch reinforcement learning such as education and healthcare. In such settings, however, unobserved variables confound observed a ctions, rendering exact evaluation of new policies impossible, i.e, unidentifiab le. We develop a robust approach that estimates sharp bounds on the (unidentifia ble) value of a given policy in an infinite-horizon problem given data from anot her policy with unobserved confounding, subject to a sensitivity model. We consi der stationary unobserved confounding and compute bounds by optimizing over the set of all stationary state-occupancy ratios that agree with a new partially ide ntified estimating equation and the sensitivity model. We prove convergence to t he sharp bounds as we collect more confounded data. Although checking set member ship is a linear program, the support function is given by a difficult nonconvex optimization problem. We develop approximations based on nonconvex projected gr adient descent and demonstrate the resulting bounds empirically.
**********

Model Class Reliance for Random Forests
Gavin Smith, Roberto Mansilla, James Goulding
Variable Importance (VI) has traditionally been cast as the process of estimatin g each variables contribution to a predictive model's overall performance. Analy sis of a single model instance, however, guarantees no insight into a variables relevance to underlying generative processes. Recent research has sought to addr ess this concern via analysis of Rashomon sets - sets of alternative model insta nces that exhibit equivalent predictive performance to some reference model, but which take different functional forms. Measures such as Model Class Reliance (M CR) have been proposed, that are computed against Rashomon sets, in order to asc ertain how much a variable must be relied on to make robust predictions, or whet her alternatives exist. If MCR range is tight, we have no choice but to use a va riable; if range is high then there exists competing, perhaps fairer models, tha t provide alternative explanations of the phenomena being examined. Applications are wide, from enabling construction of `fairer' models in areas such as recidi vism, health analytics and ethical marketing. Tractable estimation of MCR for no n-linear models is currently restricted to Kernel Regression under squared loss \cite{fisher2019all}. In this paper we introduce a new technique that extends co mputation of Model Class Reliance (MCR) to Random Forest classifiers and regress ors. The proposed approach addresses a number of open research questions, and in contrast to prior Kernel SVM MCR estimation, runs in linearithmic rather than p olynomial time. Taking a fundamentally different approach to previous work, we p rovide a solution for this important model class, identifying situations where i rrelevant covariates do not improve predictions.
**********

Follow the Perturbed Leader: Optimism and Fast Parallel Algorithms for Smooth Minimax Games
Arun Suggala, 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.
**********

Agnostic $Q$-learning with Function Approximation in Deterministic Systems: Near-Optimal Bounds on Approximation Error and Sample Complexity
Simon S. Du, Jason D. Lee, Gaurav Mahajan, Ruosong 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.
**********

Learning to Adapt to Evolving Domains
Hong Liu, Mingsheng Long, Jianmin Wang, Yu Wang
Domain adaptation aims at knowledge transfer from a labeled source domain to an unlabeled target domain. Current domain adaptation methods have made substantial advances in adapting discrete domains. However, this can be unrealistic in real-world applications, where target data usually comes in an online and continually evolving manner as small batches, posing challenges to classic domain adaptation paradigm: (1) Mainstream domain adaptation methods are tailored to stationary target domains, and can fail in non-stationary environments. (2) Since the target data arrive online, the agent should also maintain competence on previous target domains, i.e. to adapt without forgetting. To tackle these challenges, we propose a meta-adaptation framework which enables the learner to adapt to continually evolving target domain without catastrophic forgetting. Our framework comprises of two components: a meta-objective of learning representations to adapt to evolving domains, enabling meta-learning for unsupervised domain adaptation; and a meta-adapter for learning to adapt without forgetting, reserving knowledge from previous target data. Experiments validate the effectiveness our method on evolving target domains.
**********

Synthesizing Tasks for Block-based Programming
Umair Ahmed, Maria Christakis, Aleksandr Efremov, Nigel Fernandez, Ahana Ghosh, Abhik Roychoudhury, 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.
**********

Scalable Belief Propagation via Relaxed Scheduling
Vitalii Aksenov, Dan Alistarh, Janne H. Korhonen
Requests for name changes in the electronic proceedings will be accepted with no questions asked. However name changes may cause bibliographic tracking issues. Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

Firefly Neural Architecture Descent: a General Approach for Growing Neural Networks
Lemeng Wu, Bo Liu, Peter Stone, Qiang Liu
We propose firefly neural architecture descent, a general framework for progressively and dynamically growing neural networks to jointly optimize the networks' parameters and architectures. Our method works in a steepest descent fashion, which iteratively finds the best network within a functional neighborhood of the original network that includes a diverse set of candidate network structures. By using Taylor approximation, the optimal network structure in the neighborhood can be found with a greedy selection procedure. We show that firefly descent can f

lexibly grow networks both wider and deeper, and can be applied to learn accurat
e but resource-efficient neural architectures that avoid catastrophic forgetting
 in continual learning. Empirically, firefly descent achieves promising results
on both neural architecture search and continual learning. In particular, on a c
hallenging continual image classification task, it learns networks that are smal
ler in size but have higher average accuracy than those learned by the state-of-
the-art methods.
**********

Risk-Sensitive Reinforcement Learning: Near-Optimal Risk-Sample Tradeoff in Regr
et

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, Qiaomin Xie

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

Learning to Decode: Reinforcement Learning for Decoding of Sparse Graph-Based Ch
annel Codes

Salman Habib, Allison Beemer, Joerg Kliewer

We show in this work that reinforcement learning can be successfully applied to
decoding short to moderate length sparse graph-based channel codes. Specifically
, we focus on low-density parity check (LDPC) codes, which for example have been
 standardized in the context of 5G cellular communication systems due to their e
xcellent error correcting performance. These codes are typically decoded via bel
ief propagation iterative decoding on the corresponding bipartite (Tanner) graph
 of the code via flooding, i.e., all check and variable nodes in the Tanner grap
h are updated at once. In contrast, in this paper we utilize a sequential update
 policy which selects the optimum check node (CN) scheduling in order to improve
 decoding performance. In particular, we model the CN update process as a multi-
armed bandit process with dependent arms and employ a Q-learning scheme for opti
mizing the CN scheduling policy. In order to reduce the learning complexity, we
propose a novel graph-induced CN clustering approach to partition the state spac
e in such a way that dependencies between clusters are minimized. Our results sh
ow that compared to other decoding approaches from the literature, the proposed
reinforcement learning scheme not only significantly improves the decoding perfo
rmance, but also reduces the decoding complexity dramatically once
the scheduling policy is learned.
**********

Faster DBSCAN via subsampled similarity queries

Heinrich Jiang, Jennifer Jang, Jakub Lacki

Requests for name changes in the electronic proceedings will be accepted with 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.
**********

De-Anonymizing Text by Fingerprinting Language Generation

Zhen Sun, Roei Schuster, Vitaly Shmatikov

Components of machine learning systems are not (yet) perceived as security hotsp
ots.  Secure coding practices, such as ensuring that no execution paths depend o
n confidential inputs, have not yet been adopted by ML developers.  We initiate
the study of code security of ML systems by investigating how nucleus sampling--
-a popular approach for generating text, used for applications such as auto-comp
letion---unwittingly leaks texts typed by users. Our main result is that the ser
ies of nucleus sizes for many natural English word sequences is a unique fingerp
rint.  We then show how an attacker can infer typed text by measuring these fing
erprints via a suitable side channel (e.g., cache access times), explain how thi
s attack could help de-anonymize anonymous texts, and discuss defenses.
**********

Multiparameter Persistence Image for Topological Machine Learning

Mathieu Carrie■re, Andrew Blumberg

In the last decade, there has been increasing interest in topological data analysis, a new methodology for using geometric structures in data for inference and learning. A central theme in the area is the idea of persistence, which in its most basic form studies how measures of shape change as a scale parameter varies. There are now a number of frameworks that support statistics and machine learning in this context. However, in many applications there are several different parameters one might wish to vary: for example, scale and density.  In contrast to the one-parameter setting, techniques for applying statistics and machine learning in the setting of multiparameter persistence are not well understood due to the lack of a concise representation of the results.
**********

PLANS: Neuro-Symbolic Program Learning from Videos
Raphaël Dang-Nhu
Recent years have seen the rise of statistical program learning based on neural models as an alternative to traditional rule-based systems for programming by example. Rule-based approaches offer correctness guarantees in an unsupervised way as they inherently capture logical rules, while neural models are more realistically scalable to raw, high-dimensional input, and provide resistance to noisy I/O specifications. We introduce PLANS (Program LeArning from Neurally inferred Specifications), a hybrid model for program synthesis from visual observations that gets the best of both worlds, relying on (i) a neural architecture trained to extract abstract, high-level information from each raw individual input (ii) a rule-based system using the extracted information as I/O specifications to synthesize a program capturing the different observations. In order to address the key challenge of making PLANS resistant to noise in the network's output, we introduce a dynamic filtering algorithm for I/O specifications based on selective classification techniques. We obtain state-of-the-art performance at program synthesis from diverse demonstration videos in the Karel and ViZDoom environments, while requiring no ground-truth program for training.
**********

Matrix Inference and Estimation in Multi-Layer Models
Parthe Pandit, Mojtaba Sahraee Ardakan, Sundeep Rangan, Philip Schniter, Alyson K. Fletcher
Requests for name changes in the electronic proceedings will be accepted with no questions asked.  However name changes may cause bibliographic tracking issues.  Authors are asked to consider this carefully and discuss it with their co-authors prior to requesting a name change in the electronic proceedings.
**********

MeshSDF: Differentiable Iso-Surface Extraction
Edoardo Remelli, Artem Lukoianov, Stephan Richter, Benoit Guillard, Timur Bagautdinov, Pierre Baque, Pascal Fua
Geometric Deep Learning has recently made striking progress with the advent of continuous Deep Implicit Fields. They allow for detailed modeling of watertight surfaces of arbitrary topology while not relying on a 3D Euclidean grid, resulting in a learnable parameterization that is not limited in resolution.
**********

Variational Interaction Information Maximization for Cross-domain Disentanglement
HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, Kee-Eung Kim
Cross-domain disentanglement is the problem of learning representations partitioned into domain-invariant and domain-specific representations, which is a key to successful domain transfer or measuring semantic distance between two domains. Grounded in information theory, we cast the simultaneous learning of domain-invariant and domain-specific representations as a joint objective of multiple information constraints, which does not require adversarial training or gradient reversal layers. We derive a tractable bound of the objective and propose a generative model named Interaction Information Auto-Encoder (IIAE). Our approach reveals insights on the desirable representation for cross-domain disentanglement and i

ts connection to Variational Auto-Encoder (VAE). We demonstrate the validity of our model in the image-to-image translation and the cross-domain retrieval tasks. We further show that our model achieves the state-of-the-art performance in the zero-shot sketch based image retrieval task, even without external knowledge.
**********

Provably Efficient Exploration for Reinforcement Learning Using Unsupervised Learning
Fei Feng, Ruosong Wang, Wotao Yin, Simon S. Du, Lin Yang
Motivated by the prevailing paradigm of using unsupervised learning for efficient exploration in reinforcement learning (RL) problems [tang2017exploration,bellemare2016unifying], we investigate when this paradigm is provably efficient. We study episodic Markov decision processes with rich observations generated from a small number of latent states. We present a general algorithmic framework that is built upon two components: an unsupervised learning algorithm and a no-regret tabular RL algorithm. Theoretically, we prove that as long as the unsupervised learning algorithm enjoys a polynomial sample complexity guarantee, we can find a near-optimal policy with sample complexity polynomial in the number of latent states, which is significantly smaller than the number of observations. Empirically, we instantiate our framework on a class of hard exploration problems to demonstrate the practicality of our theory.
**********

Faithful Embeddings for Knowledge Base Queries
Haitian Sun, Andrew Arnold, Tania Bedrax Weiss, Fernando Pereira, William W. Cohen
The deductive closure of an ideal knowledge base (KB) contains exactly the logical queries that the KB can answer. However, in practice KBs are both incomplete and over-specified, failing to answer some queries that have real-world answers. \emph{Query embedding} (QE) techniques have been recently proposed where KB entities and KB queries are represented jointly in an embedding space, supporting relaxation and generalization in KB inference. However, experiments in this paper show that QE systems may disagree with deductive reasoning on answers that do not require generalization or relaxation. We address this problem with a novel QE method that is more faithful to deductive reasoning, and show that this leads to better performance on complex queries to incomplete KBs. Finally we show that inserting this new QE module into a neural question-answering system leads to substantial improvements over the state-of-the-art.
**********

Wasserstein Distances for Stereo Disparity Estimation
Divyansh Garg, Yan Wang, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, Wei-Lun Chao
Existing approaches to depth or disparity estimation output a distribution over a set of pre-defined discrete values. This leads to inaccurate results when the true depth or disparity does not match any of these values. The fact that this distribution is usually learned indirectly through a regression loss causes further problems in ambiguous regions around object boundaries. We address these issues using a new neural network architecture that is capable of outputting arbitrary depth values, and a new loss function that is derived from the Wasserstein distance between the true and the predicted distributions. We validate our approach on a variety of tasks, including stereo disparity and depth estimation, and the downstream 3D object detection. Our approach drastically reduces the error in ambiguous regions, especially around object boundaries that greatly affect the localization of objects in 3D, achieving the state-of-the-art in 3D object detection for autonomous driving.
**********

Multi-agent Trajectory Prediction with Fuzzy Query Attention
Nitin Kamra, Hao Zhu, Dweep Kumarbhai Trivedi, Ming Zhang, Yan Liu
Trajectory prediction for scenes with multiple agents and entities is a challenging problem in numerous domains such as traffic prediction, pedestrian tracking and path planning. We present a general architecture to address this challenge which models the crucial inductive biases of motion, namely, inertia, relative mo

tion, intents and interactions. Specifically, we propose a relational model to f
lexibly model interactions between agents in diverse environments. Since it is w
ell-known that human decision making is fuzzy by nature, at the core of our mode
l lies a novel attention mechanism which models interactions by making continuou
s-valued (fuzzy) decisions and learning the corresponding responses. Our archite
cture demonstrates significant performance gains over existing state-of-the-art
predictive models in diverse domains such as human crowd trajectories, US freewa
y traffic, NBA sports data and physics datasets. We also present ablations and a
ugmentations to understand the decision-making process and the source of gains i
n our model.
**********

Multilabel Classification by Hierarchical Partitioning and Data-dependent Groupi
ng
Shashanka Ubaru, Sanjeeb Dash, Arya Mazumdar, Oktay Gunluk
In modern multilabel classification problems, each data instance belongs to
a small number of classes among a large set of classes. In other words, these
problems involve learning very sparse binary label vectors. Moreover, in the la
rge-scale problems, the labels typically have certain (unknown) hierarchy. In th
is paper we exploit the sparsity of label vectors and the hierarchical structure
to embed them in low-dimensional space using label groupings. Consequently, we
solve the classification problem in a much lower dimensional space and then obta
in labels in the original space using an appropriately defined lifting. Our meth
od builds on the work of (Ubaru & Mazumdar, 2017), where the idea of group testi
ng was also explored for multilabel classification. We first present a novel dat
a-dependent grouping approach, where we use a group construction based on a low
-rank Nonnegative Matrix Factorization (NMF) of the label matrix of training ins
tances.
The construction also allows us, using recent results, to develop a fast predic
tion algorithm that has a \emph{logarithmic runtime in the number of labels}. W
e then present a hierarchical partitioning approach that exploits the label hier
archy in large-scale problems to divide the large label space into smaller sub-p
roblems, which can then be solved independently via the grouping approach. Numer
ical results on many benchmark datasets illustrate that, compared to other popul
ar methods, our proposed methods achieve comparable accuracy with significantly
lower computational costs.
**********

An Analysis of SVD for Deep Rotation Estimation
Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin
Rostamizadeh, Ameesh Makadia
Symmetric orthogonalization via SVD, and closely related procedures, are well-kn
own techniques for projecting matrices onto O(n) or SO(n). These tools have long
been used for applications in computer vision, for example optimal 3D alignment
problems solved by orthogonal Procrustes, rotation averaging, or Essential matr
ix decomposition. Despite its utility in different settings, SVD orthogonalizati
on as a procedure for producing rotation matrices is typically overlooked in dee
p learning models, where the preferences tend toward classic representations lik
e unit quaternions, Euler angles, and axis-angle, or more recently-introduced me
thods. Despite the importance of 3D rotations in computer vision and robotics, a
single universally effective representation is still missing. Here, we explore
the viability of SVD orthogonalization for 3D rotations in neural networks. We p
resent a theoretical analysis of SVD as used for projection onto the rotation gr
oup. Our extensive quantitative analysis shows simply replacing existing represe
ntations with the SVD orthogonalization procedure obtains state of the art perfo
rmance in many deep learning applications covering both supervised and unsupervi
sed training.
**********

Can the Brain Do Backpropagation? --- Exact Implementation of Backpropagation in
Predictive Coding Networks
Yuhang Song, Thomas Lukasiewicz, Zhenghua Xu, Rafal Bogacz
Backpropagation (BP) has been the most successful algorithm used to train artifi

cial neural networks. However, there are several gaps between BP and learning in biologically plausible neuronal networks of the brain (learning in the brain, or simply BL, for short), in particular, (1) it has been unclear to date, if BP can be implemented exactly via BL, (2) there is a lack of local plasticity in BP, i.e., weight updates require information that is not locally available, while BL utilizes only locally available information, and (3)~there is a lack of autonomy in BP, i.e., some external control over the neural network is required (e.g., switching between prediction and learning stages requires changes to dynamics and synaptic plasticity rules), while BL works fully autonomously. Bridging such gaps, i.e., understanding how BP can be approximated by BL, has been of major interest in both neuroscience and machine learning. Despite tremendous efforts, however, no previous model has bridged the gaps at a degree of demonstrating an equivalence to BP, instead, only approximations to BP have been shown. Here, we present for the first time a framework within BL that bridges the above crucial gaps. We propose a BL model that (1) produces \emph{exactly the same} updates of the neural weights as~BP, while (2)~employing local plasticity, i.e., all neurons perform only local computations, done simultaneously. We then modify it to an alternative BL model that (3) also works fully autonomously. Overall, our work provides important evidence for the debate on the long-disputed question whether the brain can perform~BP.

**********

Manifold GPLVMs for discovering non-Euclidean latent structure in neural data
Kristopher Jensen, Ta-Chu Kao, Marco Tripodi, Guillaume Hennequin

A common problem in neuroscience is to elucidate the collective neural representations of behaviorally important variables such as head direction, spatial location, upcoming movements, or mental spatial transformations. Often, these latent variables are internal constructs not directly accessible to the experimenter. Here, we propose a new probabilistic latent variable model to simultaneously identify the latent state and the way each neuron contributes to its representation in an unsupervised way. In contrast to previous models which assume Euclidean latent spaces, we embrace the fact that latent states often belong to symmetric manifolds such as spheres, tori, or rotation groups of various dimensions. We therefore propose the manifold Gaussian process latent variable model (mGPLVM), where neural responses arise from (i) a shared latent variable living on a specific manifold, and (ii) a set of non-parametric tuning curves determining how each neuron contributes to the representation. Cross-validated comparisons of models with different topologies can be used to distinguish between candidate manifolds, and variational inference enables quantification of uncertainty. We demonstrate the validity of the approach on several synthetic datasets, as well as on calcium recordings from the ellipsoid body of Drosophila melanogaster and extracellular recordings from the mouse anterodorsal thalamic nucleus. These circuits are both known to encode head direction, and mGPLVM correctly recovers the ring topology expected from neural populations representing a single angular variable.

**********

Distributed Distillation for On-Device Learning
Ilai Bistritz, Ariana Mann, Nicholas Bambos

On-device learning promises collaborative training of machine learning models across edge devices without the sharing of user data. In state-of-the-art on-device learning algorithms, devices communicate their model weights over a decentralized communication network. Transmitting model weights requires huge communication overhead and means only devices with identical model architectures can be included. To overcome these limitations, we introduce a distributed distillation algorithm where devices communicate and learn from soft-decision (softmax) outputs, which are inherently architecture-agnostic and scale only with the number of classes. The communicated soft-decisions are each model's outputs on a public, unlabeled reference dataset, which serves as a common vocabulary between devices. We prove that our algorithm converges with probability 1 to a stationary point where all devices in the communication network distill the entire network's knowledge on the reference data, regardless of their local connections. Our analysis assumes smooth loss functions, which can be non-convex. Simulations support our t

heoretical findings and show that even a naive implementation of our algorithm s ignificantly reduces the communication overhead while achieving an overall compa rable performance to state-of-the-art, depending on the regime. By requiring lit tle communication overhead and allowing for cross-architecture training, we remo ve two main obstacles to scaling on-device learning.
**********

COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learnin g

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, Thomas Brox

Many real-world video-text tasks involve different levels of granularity, such a s frames and words, clip and sentences or videos and paragraphs, each with disti nct semantics. In this paper, we propose a Cooperative hierarchical Transformer (COOT) to leverage this hierarchy information and model the interactions between different levels of granularity and different modalities. The method consists o f three major components: an attention-aware feature aggregation layer, which le verages the local temporal context (intra-level, e.g., within a clip), a context ual transformer to learn the interactions between low-level and high-level seman tics (inter-level, e.g. clip-video, sentence-paragraph), and a cross-modal cycle -consistency loss to connect video and text. The resulting method compares favor ably to the state of the art on several benchmarks while having few parameters.
**********

Passport-aware Normalization for Deep Model Protection

Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Gang Hua, Nenghai Yu

Despite tremendous success in many application scenarios, deep learning faces se rious intellectual property (IP) infringement threats. Considering the cost of d esigning and training a good model, infringements will significantly infringe th e interests of the original model owner. Recently, many impressive works have em erged for deep model IP protection. However, they either are vulnerable to ambig uity attacks, or require changes in the target network structure by replacing it s original normalization layers and hence cause significant performance drops. T o this end, we propose a new passport-aware normalization formulation, which is generally applicable to most existing normalization layers and only needs to add another passport-aware branch for IP protection. This new branch is jointly tra ined with the target model but discarded in the inference stage. Therefore it ca uses no structure change in the target model. Only when the model IP is suspecte d to be stolen by someone, the private passport-aware branch is added back for o wnership verification. Through extensive experiments, we verify its effectivenes s in both image and 3D point recognition models. It is demonstrated to be robust not only to common attack techniques like fine-tuning and model compression, bu t also to ambiguity attacks. By further combining it with trigger-set based meth ods, both black-box and white-box verification can be achieved for enhanced secu rity of deep learning models deployed in real systems.
**********

Sampling-Decomposable Generative Adversarial Recommender

Binbin Jin, Defu Lian, Zheng Liu, Qi Liu, Jianhui Ma, Xing Xie, Enhong Chen

Recommendation techniques are important approaches for alleviating information o verload. Being often trained on implicit user feedback, many recommenders suffer from the sparsity challenge due to the lack of explicitly negative samples. The GAN-style recommenders (i.e., IRGAN) addresses the challenge by learning a gene rator and a discriminator adversarially, such that the generator produces increa singly difficult samples for the discriminator to accelerate optimizing the disc rimination objective. However, producing samples from the generator is very time -consuming, and our empirical study shows that the discriminator performs poor i n top-k item recommendation. To this end, a theoretical analysis is made for the GAN-style algorithms, showing that the generator of limit capacity is diverged from the optimal generator. This may interpret the limitation of discriminator's performance. Based on these findings, we propose a Sampling-Decomposable Genera tive Adversarial Recommender (SD-GAR). In the framework, the divergence between some generator and the optimum is compensated by self-normalized importance samp ling; the efficiency of sample generation is improved with a sampling-decomposab

le generator, such that each sample can be generated in O(1) with the Vose-Alias method. Interestingly, due to decomposability of sampling, the generator can be optimized with the closed-form solutions in an alternating manner, being different from policy gradient in the GAN-style algorithms. We extensively evaluate the proposed algorithm with five real-world recommendation datasets. The results show that SD-GAR outperforms IRGAN by 12.4% and the SOTA recommender by 10% on average. Moreover, discriminator training can be 20x faster on the dataset with more than 120K items.

**********

Limits to Depth Efficiencies of Self-Attention

Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, Amnon Shashua

Self-attention architectures, which are rapidly pushing the frontier in natural language processing, demonstrate a surprising depth-inefficient behavior: Empirical signals indicate that increasing the internal representation (network width) is just as useful as increasing the number of self-attention layers (network depth). In this paper, we theoretically study the interplay between depth and width in self-attention. We shed light on the root of the above phenomenon, and establish two distinct parameter regimes of depth efficiency and inefficiency in self-attention. We invalidate the seemingly plausible hypothesis by which widening is as effective as deepening for self-attention, and show that in fact stacking self-attention layers is so effective that it quickly saturates a capacity of the network width. Specifically, we pinpoint a ``depth threshold" that is logarithmic in the network width: for networks of depth that is below the threshold, we establish a double-exponential depth-efficiency of the self-attention operation, while for depths over the threshold we show that depth-inefficiency kicks in. Our predictions accord with existing empirical ablations, and we further demonstrate the two depth-(in)efficiency regimes experimentally for common network depths of 6, 12, and 24. By identifying network width as a limiting factor, our analysis indicates that solutions for dramatically increasing the width can facilitate the next leap in self-attention expressivity.

**********